University of Szeged
Graduate School in Linguistics
English Applied Linguistics PhD Program

Establishing the Context and Scoring Validity of the Writing Tasks of Euroexam International's English for Academic Purposes Test

PhD Dissertation

Fűköh Borbála

Supervisor: Dr Barát Erzsébet

Szeged
2020

Acknowledgements

First and foremost I would like to thank Euroexam International for supporting my research. I am indebted to Zoltán Rozgonyi and Kristóf Hegedűs who granted me permission to collect and use Euroexam data; István Török and Judit Sutherland who helped me compile exam tasks and other documentation. I am thankful to all the participants of my research for their cooperation and dedication.

I acknowledge the invaluable advice of my colleague, Zoltán Lukácsi throughout the research stages.

I wish to thank my supervisor, Erzsébet Barát for her competent guidance and support through the entire length of my PhD studies.

Finally, I am grateful to my family for their patience and support.

Abstract

The main objective of the dissertation is to develop validity arguments in support of the writing tasks of a C1 level English for Academic Purposes (EAP) test to be implemented by Euroexam International, Budapest. I deployed a research-based development process to build a validity argument about how the construct of the proposed writing tasks in an Academic test reflects the skills required in higher education, and whether the results reflect reliable scores and unbiased marking. The dissertation focuses on the two proposed writing tasks of the EAP test (*formal transactional email* and *discussion essay*). The research covers (a) the initial development stage; (b) the completion of the specifications and the test items; (c) the trialling and pre-testing of test items. Since the evidencing of objective and unbiased marking is one major requirement expected by international tertiary education institutions, the dissertation also aimed to collect and analyse data to improve the scoring validity of the essay task; and proposed a level a genre specific checklist-based rating tool instead of the C1 level accredited rating scale of Euroexam International. The method is built upon Weir's (2005a) theoretical framework and the characteristics of test usefulness (Bachman & Palmer, 1996; 2010), and consider Read's (2015) validation stages, using a mixed-methods approach.

Table of Contents

Chapter 1: Introduction

High-stakes language testing plays an important role in the education system of Hungary. The language exams are supervised by the Accreditation Board for Foreign Language Examinations, which is responsible for standardising the professional requirements for examination boards across the country. Currently both major international test providers and locally developed exams are available for test takers (Educational Authority, 2020b).

The major test providers, such as Pearson, IELTS and TOEFL, offer academic tests for students who desire to continue their studies in English language higher education (IELTS, 2018; Pearson PTE Academic, 2017; TOEFL iBT, 2018). Despite the growing number of Hungarian students pursuing university studies in European Union and UK universities, there are no state accredited English for Academic Purposes (EAP) exams available in Hungary yet. In 2017, Euroexam International decided to launch an exam development project to make up for this gap and designed an English for Academic Purposes (EAP) test targeted at Hungarian and East-Central European students (Euroexam Academic, 2019a). As member of the Euroexam Research Team as well as an ESP instructor in tertiary education, I undertook the task of leading the validation research for the writing tasks of the test (Fűköh, 2019a, 2019b). Validity evidence is necessary when designing a new test in order to determine what tasks reflect best the skills actually required in higher education (Bachman & Palmer, 1996; Read, 2015; Weir, 2005a).

My dissertation aims to develop validity arguments in support of the writing tasks of a C1 level English for Academic Purposes (EAP) test to be implemented by Euroexam International, Budapest in 2019. For test development, I deployed a research-based development process to see whether and to what extent the task reflects the skills required in higher education, and whether the results reflect reliable scores and unbiased marking. The aim of this research is therefore to establish the validity of the two proposed writing tasks (*formal transactional email* and *discussion essay*). The research was designed to comprise (a) the development stage; (b) the completion of the specifications and the test items; (c) the trialling and pre-testing of test items; and (d) aims to collect and analyse data to establish the scoring validity of checklist-based marking for the discussion essay. The methodology of generating validity evidence adopts Weir's (2005) validation framework that draws on a mixed-methods approach, and it is rendered into the stages proposed by Read (2015).

The research is of relevance for language teachers and all the stakeholders of a high-stakes language test (Yin, 2011, pp. 73-74) through producing evidence-based writing tasks and grading protocols. The stages of validation are to prove that the writing tasks of the Academic Exam of Euroexam International reflect the skills needed in the different academic discourse types students need to perform in the course of their learning.

The dissertation is divided into two main parts. The first part of the dissertation is a review of the relevant literature in four chapters. After introducing the background to the research, Chapter 2 brings together the relevant theoretical works on the nature of writing and the writing process with a special focus on the nature of academic writing.

Chapter 3 is devoted to the topic of assessing writing. In addition to discussing the nature of writing assessment in general, and second language writing as it appears in the Common European Framework of Reference (Council of Europe, 2001), I focus on raters, rater leniency and harshness, and rater training. The last part of the chapter explores the use of rating scales and checklists, their advantages and disadvantages.

In Chapter 4, I give an overview of test development, more specifically test validity. I discuss the development of language tests, more particularly the characteristics of test usefulness as presented by Bachman and Palmer (1996; 2010). The second part of this chapter is devoted to the various concepts of validity in language testing. I present the traditional and the new concepts of validity and discuss the socio-cognitive framework by Weir (2005a). In addition, I discuss the concept of localisation that is of particular relevance in the context of international university admissions.

Chapter 5 is focused on the methods of generating validity evidence and argue for the advantage of mixed-methods research in the test development process. This is the chapter that formulates and argues for the research questions of the dissertation in the context of my main area of research.

I present my empirical research in the second part of the dissertation in three chapters. These chapters follow the stages in the test development process of the Euroexam Academic project. Chapter 6 outlines the initial development phase, in other words, the planning and domain analysis phases of the test development process as well as the reflection on the judgement of an external expert panel.

Chapter 7 presents two stages of the validation research: trialling and pretesting. Before completing the test specifications, a detailed description of the test and a trial

version of the test tasks are compiled and test taker and rater feedback is collected through semi-structured interviews. Chapter 7 first presents the qualitative data collection and analysis of test taker and task characteristics based on a blueprint of the specifications (Read, 2015, p. 177). The trialling of the writing tasks of the test was conducted with a small sample of test takers and Euroexam raters. I present the textual analysis of the qualitative data sets of test taker performance and rater think aloud rating processes. After the qualitative data collection and analysis in the course of a small-scale trial, I present the collection of quantitative data in the course of pretesting the proposed test tasks. As for pretesting, the recommended test development protocol was followed: the test paper was administered with a pretest population which was similar to the target population of the academic test; the result of the pretest was analysed using Classical Test Theory.

The findings of the qualitative and quantitative data analysis have twofold relevance. On the one hand, the analysis of the verbal protocols and the large-scale pretesting help establish the validity of the writing tasks. On the other hand, they shed light on the issues of scoring validity I problematized in Chapter 8. In this chapter, I revisit the issues identified with rating in the previous stages and discuss the development of a genre and level specific checklist-based rating tool developed for ensuring the objective and unbiased nature of the rating procedure.

Finally, the results of the thesis are summed up in the Conclusion chapter that indicates the suitability of rating on a checklist as a potential direction of the current research. Further to the focus of the current research, i.e. the checklist-based rating tool for the C1 level discussion essay, I highlight the need for the development of checklist-based rating tools for the academic genres that are detailed in the test specifications and the genres that appear in the C1 level General English test of Euroexam International.

The illustrations of the dissertation are labelled sequentially. Text or numbers in the form of columns are named Tables, whereas illustrations such as drawings, graphs or illustrations taken from other authors' work are referred to as Figures. The labelling of the illustrations and references follows the APA Referencing Style Guide.

Chapter 2: The Nature of Writing Ability

In this chapter, I deal with the understanding of the nature of writing ability, the construct of second language writing in language tests and the different aspects of writing skills. Writing ability is closely connected to academic and professional success (Weigle, 2002, p. 4), thus writing skills, even in one's first language (L1), need to be explicitly taught. Writing when compared to speaking, is clearly a more standardised system, this is one of the reasons why it must be acquired through instruction (Grabowsky, 1996, p. 75). Existing cultural differences may cause problems for second language (L2) learners, especially in the field of academic writing. A number of studies claim that success in the academic discourse community is mainly measured with the level of writing skills (Cameron, 2000; Spack, 1988; Weninger & Khan, 2013), thus language tests may have an "impact on the career or life chances of individual test takers" (Taylor, 2005, p. 154).

First, I review the writing models, then I discuss the differences between the understanding of L1 and the L2 writing with special emphasis on the nature of academic writing and its cognitive as well as social and cultural aspects.

2.1 Models of L1 writing

Experts working in language test development need to take into account different proficiency levels regarding writing skills. As early as 1980, Flower and Hayes moved away from the product-oriented approach to writing and emphasized the cognitive processes of writers (Flower & Hayes, 1980). In this section, I present different process oriented cognitive models.

Hayes and Flower (1980) in their influential work emphasize the writer's intention and their efforts made during the thinking process. They identify three components of the writing process, such as the task environment, the writer's long-term memory and the processes the writer engages in. They focus on three main sub-processes of writing: (a) planning, (b) translating, and (c) reviewing. It is important to highlight that these components and sub-processes interact with one another (as indicated by arrows in Figure 1). The writing process according to the model is *recursive* as opposed to the *linearity* of the end product models. The task environment consists of both the writing assignment and the text produced so far; with the latter keeping a constant connection with the translating and reviewing processes of the writer. Translating and reviewing together with planning

are part of the monitoring process, which belongs to the mental processes of metacognition.

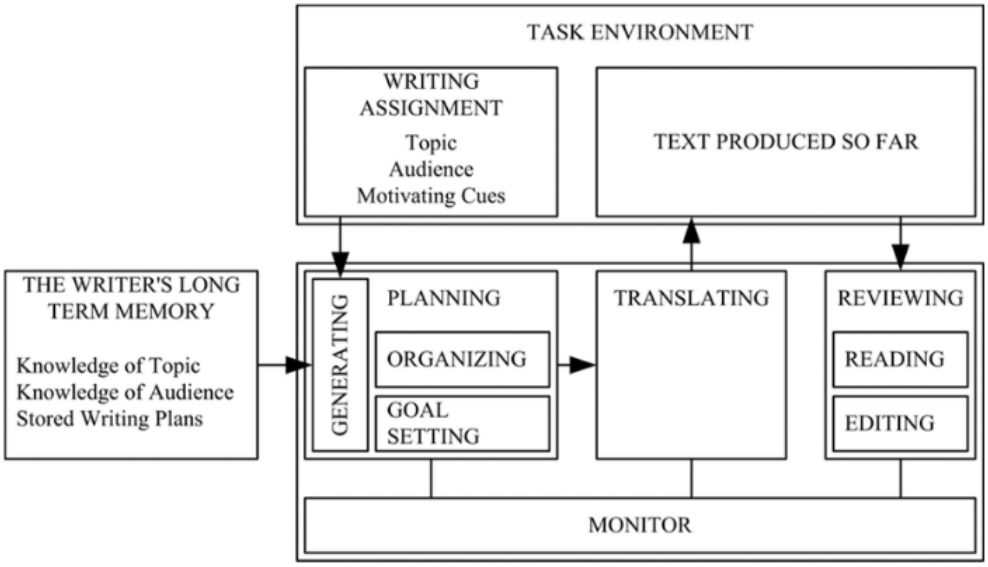


Figure 1. The Hayes and Flower model (1980) (Weigle, 2002, p. 24)

Subsequent research into the metacognitive processes by Bereiter and Scardamalia (1982) complemented the Hayes and Flower model by describing in detail what mechanisms writers use in the different stages of writing. They also emphasize the role of instruction so that writers know when to move from one process to the next and use the monitoring activity effectively. Bereiter and Scardamalia (1987) conducted further research using think aloud protocols among *novice* and *expert* writers and found that these complex mechanisms mainly support techniques used by expert writers.

Based on the differentiation between unskilled (*novice*) and skilled (*expert*) writers, Scardamalia and Bereiter (1987) presented two models: the model of knowledge telling and knowledge transforming, respectively, which have been widely used in writing assessment literature. The former process is argued to "make writing a fairly natural task", whereas the latter "makes writing a task that keeps growing in complexity to match the expanding competence of the writer" (p. 5).

As Figure 2 shows, the knowledge-telling model is rather linear. We can see that the writer uses their existing content and discourse knowledge to fulfil the writing task. Although Bereiter and Scardamalia point out that it is possible to construct a well-formed quality text using the knowledge-telling model, provided the topic and the text type required by the task are familiar to the writer, but it is clearly discernible from the model

that there is a strong reliance on memory probes instead of using the metacognitive processes.

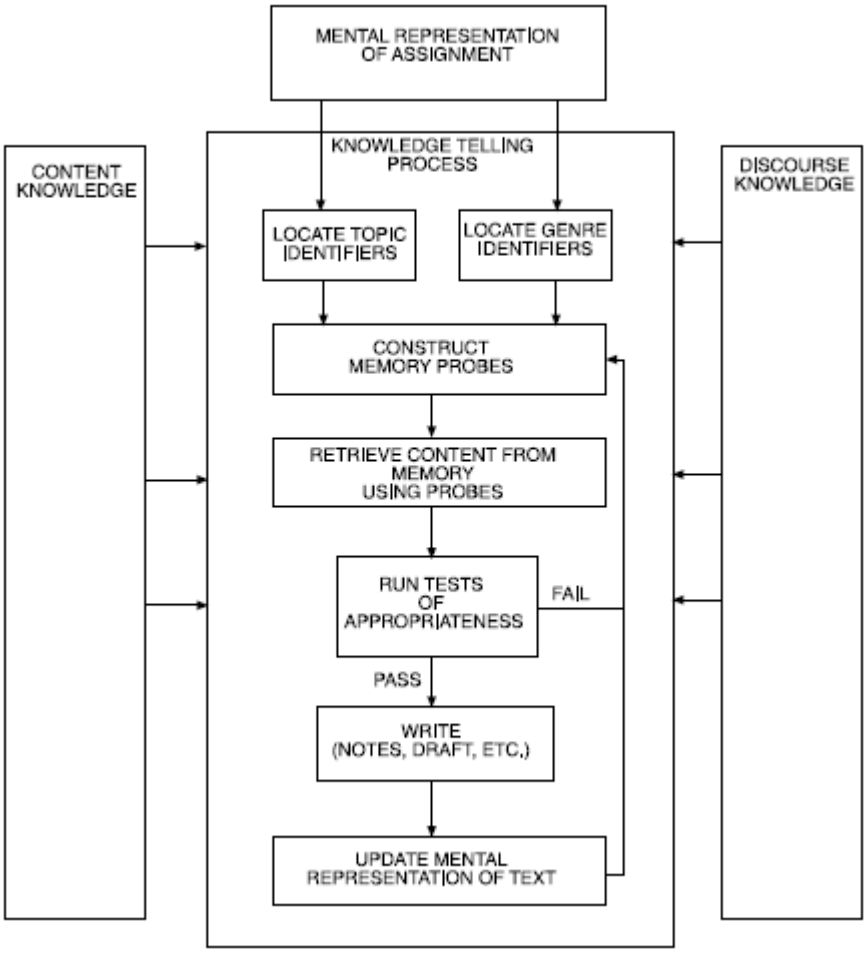


Figure 2. Structure of the knowledge-telling model (Bereiter and Scardamalia, 1987, p. 8)

The knowledge-processing model is presented in Figure 3. According to the model, skilled writers will problematize the gap between the task requirements and their resources. As Weigle (2002) notes in her assessment of the model, this results in "problem solving activities in two domains, called the content problem space and the rhetorical problem space" (p. 33). When the problems are solved, they are transferred to the knowledge-telling phase and the expert writer composes the written product. The process in real life is iterative: when new problems arise, the process goes back to the problem analysis and goal setting stage. This recursive nature of writing was not highlighted in earlier models.

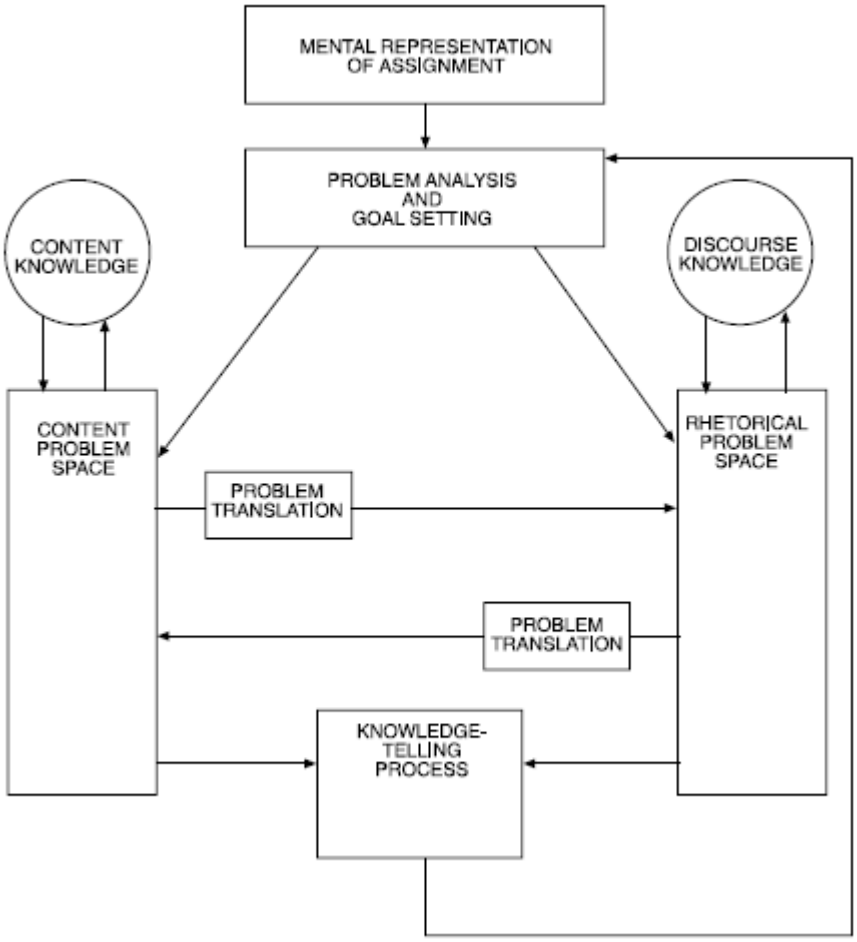


Figure 3. Structure of the knowledge-transforming model (Bereiter and Scardamalia, 1987, p. 11)

Importantly, the model accounts for the distinction between novice and expert writers, nevertheless it has its limitations. Firstly, it lacks an explanation for how novice writers may become expert writers (Grabe & Kaplan, 1996). Secondly, based on L1 writing, the model does not deal with the role of linguistic knowledge in an L2 writing context. As Di Gennaro (2006, pp. 3-5) points out, L2 writing research has hugely benefited from L1 writing models, nevertheless existing linguistic competence and culture specific textual organisation patterns were left out of their scope.

Based on the advancement of communicative language teaching, the 1990s saw a development in research in language ability and sociolinguistic theory. As regards communicative language ability, in their model, Bachman (1990) and Bachman and Palmer (1996) argued that language ability is made up of two parts: language competence and strategic competence, where the former is further divided into organisational and

pragmatic knowledge. Based on this model, in which sociolinguistic knowledge and non-linguistic factors play a crucial role, Hayes (1996) redefined and expanded the early Hayes and Flower (1980) writing model. He put the individual in the centre of the model and added two external components: the social and the physical environment. Instead of the planning, translating and reviewing processes of the earlier model, reflection, text production and text interpretation appear as the cognitive processes of the individual learner. The new model also stressed the interactive nature of the different components within the individual, and the individual's interaction with the external components (Figure 4).

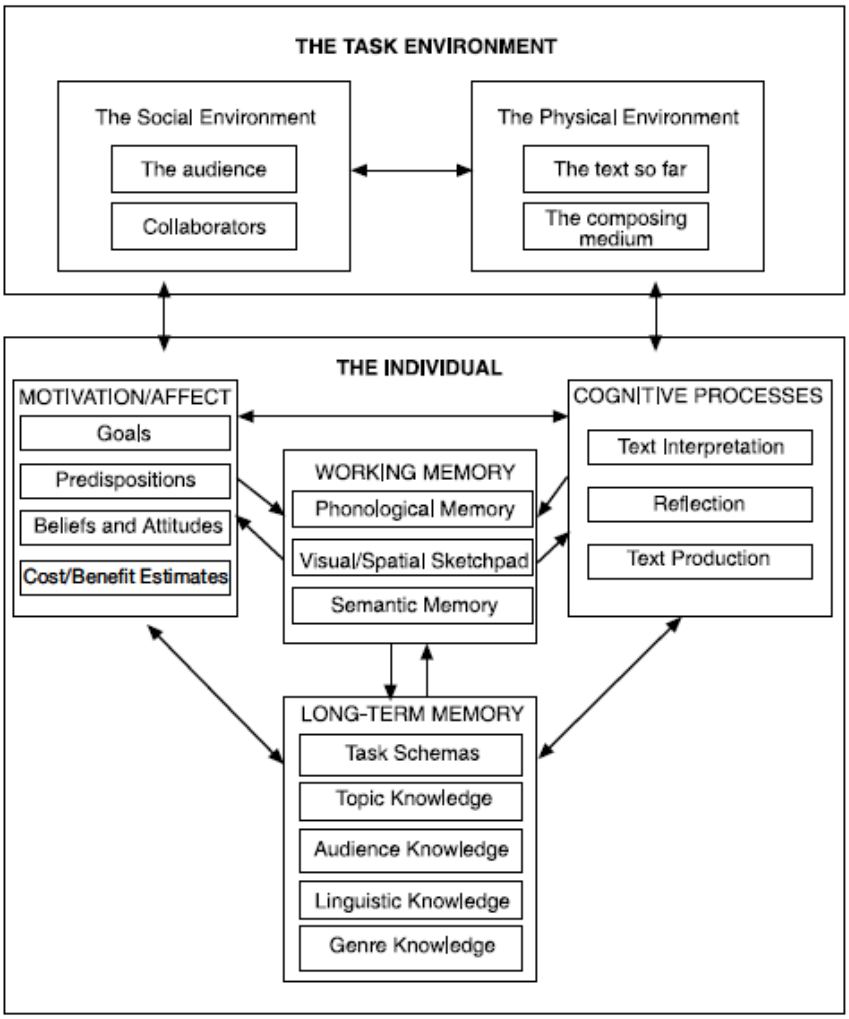


Figure 4. The revised Hayes model (Hayes, 1996, p. 10)

Despite its enhanced complexity regarding the number of internal and external components, their relationships and a more detailed description of long-term memory, the new model still has its shortcomings in terms of L2 writing, Linguistic knowledge appears as a component within the writer's long-term memory, and there is no adequate

explanation regarding the nature of interaction among the different components. As Weigle put forward, (2002, pp. 28-29), the Grabe and Kaplan-model (1996) may be regarded as a supplement that incorporates the model of communicative language use, and thus gives account for the characteristics of second language writing (see below).

2.2 Models of L2 academic writing

As concluded above in Section 2.1, no matter how influential L1 writing models are, L2 writing is different in many ways. L2 students' social and cognitive factors affect their writing styles, their culture-specific schemata are not directly transferable to foreign language writing (Myles, 2002, p. 2). Although practical handbooks of academic writing often stress that there is a "generally accepted way" (Hartley, 2008, p. 3) of writing academic and scientific texts, and that there are "transnational features" (Swales, 1990, p. 24) writers tend to follow, there are different cultural features one must take into consideration when practising or instructing academic writing. Kaplan's (1966) early work in contrastive rhetoric, for instance, stresses that L2 writers' different thinking patterns, mental processes and writing conventions should not be ignored. Furthermore, Connor (1997) highlights that L2 student writing might show the surface features of L1 academic writing, but they rarely fit into the patterns of the particular genre.

Matsuda (1997) claims that writers' native language, culture and education greatly affect the discourse patterns of their L2 text. In other words, the organisational patterns of an L2 text are much more influenced by the learner's L1 than their L2 language level. Krapels (1990), when comparing findings of L1 and L2 writing processes, points out that L1 writing processes are often transferred to L2 writing processes. In his research, Woodall (2002) found L1 resources highly important to rely on in the L2 writing process. For this reason, I argue that integrating students into the English-speaking academic community is not without difficulties. In the context of language testing, it is important to identify which parts of the writing process can be used to describe L2 writing ability so that they can be used for test development and building a validity argument.

Research into communicative language ability shed light on the need of language competence for L2 writing, which generated more interest in what constitutes L2 writing ability. Chapelle et al. (1993) proposed a model for academic language use, in which they described the four skills of communicative language use (listening, speaking, reading

writing). This model was further adapted by Grabe and Kaplan (1996) to suit L2 writing skills (Figure 5).

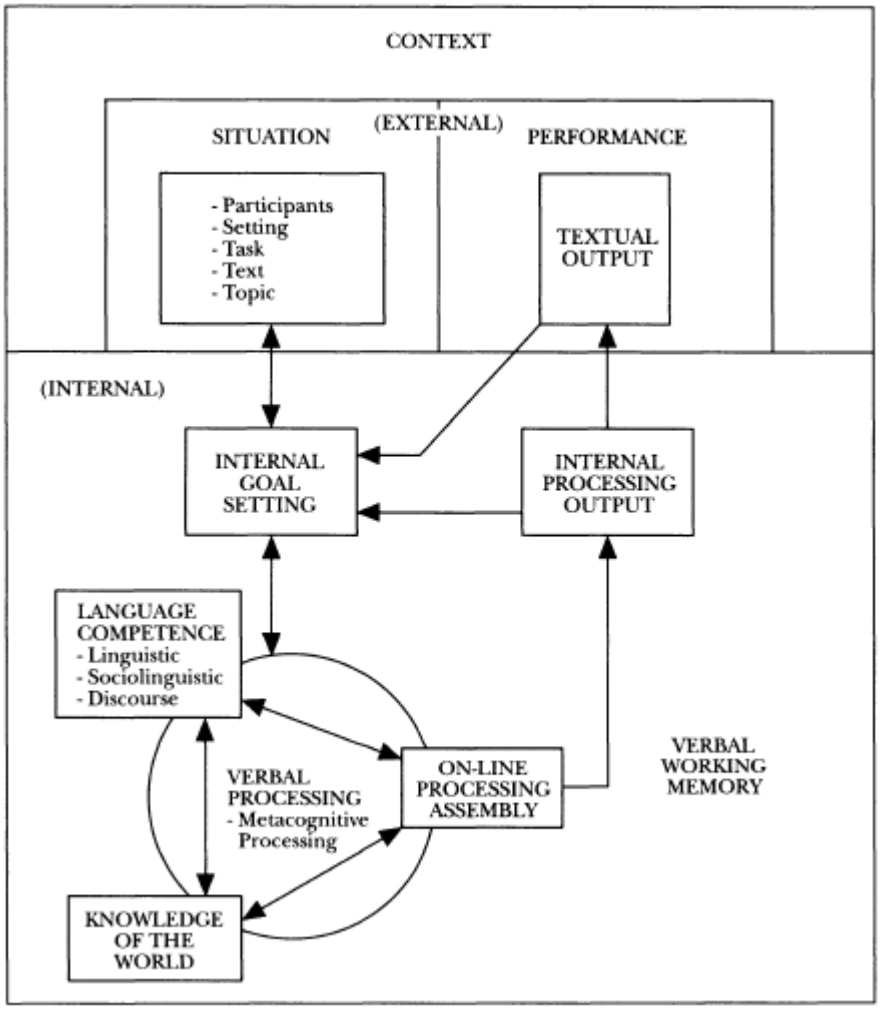


Figure 5. Model of writing as communicative language use (Grabe & Kaplan, 1996, p. 226)

With writing presented as an iterative process, the model is very different from the earlier process approaches. Its novelty lies in the

> "incorporation of textual influences, the explicit specification of the context, and the built-in comparison mechanism between the goal-setting component and the three sources of processing/processing outcomes (verbal processing, internal processing output, textual output)" (Grabe & Kaplan, 1996, p. 229).

The model (Figure 5) clearly manages in great detail to deal with language competence. Grabe and Kaplan divide language competence into three elements of knowledge: linguistic, sociolinguistic and discourse knowledge. They stress the importance of the ways

language is used in different social settings and the competence of structuring a coherent text apart from the knowledge of the linguistic elements of a language. They also provide a detailed taxonomy of language competence consisting of (a) Linguistic knowledge, (b) Discourse knowledge, and (c) Sociolinguistic knowledge (Grabe & Kaplan, 1996, pp. 220-21), which has been overtly used in the practice of assessing writing.

All the above models contain the element of context as the ground of meaningful communication. The context, however, is not only an external factor that influences language use, but the various linguistic forms themselves are rooted in social practices of language users over time. Second language acquisition (SLA) and instructed language learning have both been greatly influenced by Vygotsky's sociocultural theory (SCT) and Halliday's systemic functional linguistics (SFL) (Byrnes, 2006). A number of researchers (Ellis, 2000; Lantolf, 2000; Swain 2002) based their studies on SCT (Vygotsky, 1978) and argued that language learning is a social process and stressed the role of socio-cultural circumstances central in learners' cognitive development. Halliday (1994; Halliday & Matthiessen, 2004) in his systemic functional linguistics (SFL) also views language as a social phenomenon, and suggests that both the lexico-grammatical categories and the language users' choices are constructed under the influence of the social and cultural context. As regards writing research, Hamp-Lyons and Kroll (1997) also acknowledge the social nature of writing and claim that the social and contextual aspects of the writing process shapes the writing product so that it be appropriate for an intended audience. Sperling (1996) claims that writing is "a meaning making activity that is socially and culturally shaped and individually and socially purposeful" (p. 55).

After Kaplan's (1966) work on contrastive rhetoric, a number of studies were published to reveal the cultural differences and culture-specific patterns of discourse that characterise writing (Collado, 1981; Leki, 1992; Matalene, 1985; Ostler, 1987 all cited in Weigle, 2002). To provide a framework for the systemic investigation of the cultural aspects of writing, Matsuda (1997) suggested that contrastive rhetoric model move towards a more dynamic model (Figure 6) that contains the writer's cultural background and reflects the complexity of the process the writer goes through.

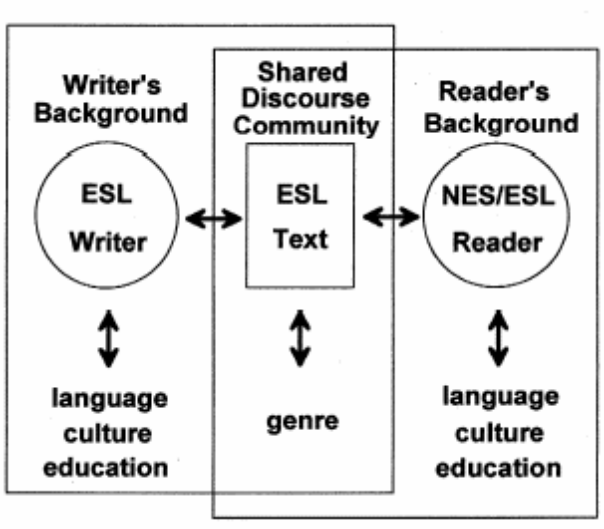


Figure 6. Matsuda's dynamic model of L2 writing (Matsuda, 1997, p. 52)

As it is displayed in Figure 6, the model is dynamic because there is a relationship between the reader and the writer mediated through their membership in a discourse community. Both the reader and the writer have their own backgrounds with their dialects, socioeconomic status, subject knowledge, but these backgrounds are complex and flexible, and come from an ever-changing shared discourse community. Although the model does not make a difference between L1 and L2 writers and readers, but it stresses the ongoing interaction between the elements. As early as the 1980's, Land and Whitely (1989) pointed out that the social reality of the writer and the reader is created through their negotiation of the context. This view greatly matches the plurilingual profile of the learner in the *Companion Volume of the Common European Framework of Reference: Learning, teaching, assessment* (Council of Europe, 2018), which complements the CEFR as the key document for language syllabuses, curriculum guidelines and language assessment. It introduced new 'pluri-' scales to recognize and support the plurilingualism and pluriculture of the social agent. It presents language use as "a dynamic, never-ending process to make meaning using different linguistic and semiotic resources" (Piccardo, 2018, p. 9). In Chapter 3, I discuss the conceptualizing of writing in the CEFR, while in Chapter 8 I explore the relevant CEFR scales in detail.

The socio-cognitive model (Shaw & Weir, 2007; Weir, 2005a) also conceptualizes writing as an interaction between the context of use and the writer's cognitive processes. The model, having been designed to account for high-stakes language test processes, acknowledges the social and cultural factors that shape the writing process, the linguistic

demands second language writers have to meet, and the personal characteristics of the test taker. Although these elements appear in earlier models, the socio-cognitive framework reconfigures them and attends to the lack of those models, identifying the different levels of cognitive processing. Shaw and Weir (2007) based their model on Field's model (2004) of information processing and upon Kellog's (1994) idea of the information processing phases. Kellog identifies a stage, which includes generating ideas, organising and setting goals (*planning*). In Field's model, planning is divided into three stages macro-planning, organisation and micro-planning. Shaw and Weir argued that writing, including academic writing, is not a linear activity bur rather recursive that consists of five processes: (a) macro-planning, (b) organisation, (c) micro-planning, (d) translation, and (e) monitoring (2007, p. 34). Moreover, they claim that writing is not an isolated activity but is in interaction with external and internal factors. The selection of individual cognitive processes and the way of putting them in action are based on the 5-element planning process.

Shaw and Weir's cognitive processing framework for writing appears as part of Weir's (2005) earlier framework of test validation. An important point he draws attention to for the assessment of writing in high-stakes language tests is to identify which phases and cognitive processes can be considered relevant for test development and validity. I shall discuss Weir's socio-cognitive framework of validation in detail in Chapter 4 in the context of test validation. To show the internal and external components and the specific features of writing, I present below the updated framework (Figure 7) of Shaw and Weir (2007). The model itself can be divided into two main parts *a priori* and *a posteriori* validation. The graphical representation makes clear how the different components interact and how they fit together conceptually.

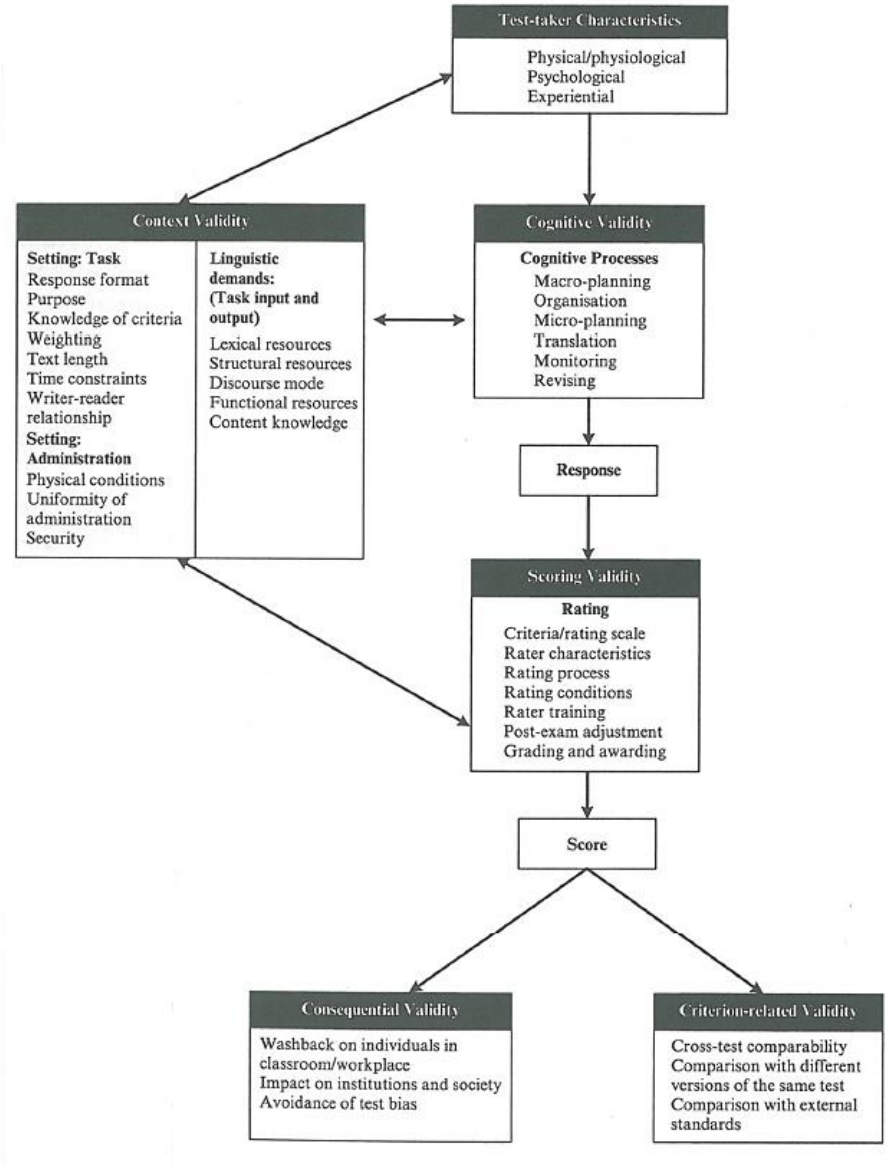


Figure 7. A framework of conceptualising writing test performance (Shaw & Weir, 2007, p. 4)

The framework is called *socio-cognitive* because the skills involved that are tested are demonstrated by the mental processing of the candidate; at the same time the writing performance is regarded as a social rather than a purely linguistic phenomenon: In addition to the context, the test-taker's individual characteristics also appear, and are in an immediate relationship with context and cognitive validity. The personal characteristics can be divided into three categories: (a) physical/physiological categories, (b) psychological characteristics, and (c) experiential characteristics. In other words, the test taker's characteristics, including age, interest, experience, knowledge and motivation may

have an impact on their performance. These factors need to be taken into consideration so that no test takers become disadvantaged in the testing process.

The other element a writing model needs to contain in order to suit the specificities of second language writing is the language skill requirement. Although some earlier models lacked this aspect, it is evident that the process of generating a written text by L2 writers is also shaped by the linguistic demands of the task. Linguistic demands in the socio-cognitive model are part of context validity, and they are closely connected to the personal characteristics. The interconnected nature of linguistic demand, cognitive processes and individual characteristics is overtly important in connection with second language writing. L2 writers might be disadvantaged compared to L1 writers by their linguistic abilities, their social and cultural background. At the same time, Weigle (2002, p. 37) points out that very little research is there dealing with the connection between writing quality and writing apprehension, while the latter may well be a greater issue for L2 writers than for L1 ones.

As specified in the Euroexam Detailed Specifications (2019b), the construct of the writing test uses the Grabe and Kaplan model (1996) as a theoretical foundation. In addition to this, the rating scales are based on the CEFR (Council of Europe, 2001) descriptive scales. In the current research-based test development project, I rely on the Euroexam writing construct; furthermore, I am adapting the design of the present research-based validation process to the components of writing test performance as presented in Shaw and Weir's validation framework (2007).

Chapter 3: Assessing Writing

After reviewing the different models of the writing process, in this chapter I focus on questions more closely related to the assessment of second language writing in the context of language testing. Assessing writing involves a two-purpose decision: making inferences about the test taker's ability and making decisions based on the inferences (Bachman & Palmer, 1996). Today, the latter is especially important since the decisions concerning writing ability in high-stakes testing influence test takers' future success in life and access to other commodities (Fairclough, 1999; Hamp-Lyons, 2003). In the first part of this chapter, I summarise the different purposes of tests and the task types used in assessing writing ability. Secondly, I review the international standard for assessing foreign language abilities in the *European Union, Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001; 2018) and its use for assessment purposes. My ultimate aim is to focus on the limitations of the framework in the context of assessing writing products.

As the assessment of foreign language writing is typically viewed to be a subjectively performed task in language testing, I discuss the relevant theoretical works, including the main issues involved in the rating process and the different stages and conditions of the process. I will look at the questions of rater-mediated assessment (McNamara, 2000, p. 34) and that of rating training, rater feedback research and quality control procedures.

I will conclude the chapter with the evaluation of the suitability of two approaches to assessing writing products. I discuss two different rating tools for subjectively marked tasks: rating scales and checklists in relation to multilevel and level specific tests.

3.1 The fundamentals of assessing writing

Cronbach (1984) argues that a test is always chosen for a particular situation and has a concrete purpose. The purpose usually is to make inferences about language ability and sometimes further decisions are made based on the inferences. These inferences and decisions made by the assessors are valid in the particular context. As for the inferences, language tests can be divided into three groups: (a) proficiency, (b) diagnostic and (c) achievement ones. The decisions made based on the inferences can be low-stakes or high-stakes (Weigle, 2002, pp. 40-41). Low stakes tests have low, high-stakes tests have high impact on test takers' lives.

In order to make inferences and decisions based on tests, we must define what to assess and how to assess it. Ready-made tests rarely serve the purpose, this is the reason why teachers are advised, and examination boards are required to develop their own tests (Fulcher, 2010, p. 101). The process of test development is discussed in detail in Chapter 4. Here I discuss the concepts behind assessment: the primary aim of language testing and the considerations for the task types to be used.

The most important component in language testing is the ability we would like to test. This is ability is referred to as the construct. Alderson (2000, p. 118) claims that constructs are abstractions that exist in particular definitions of assessment purpose, which means that it is impossible to give a definition for language that is applicable in all situations. Furthermore, as it was highlighted in Chapter 2, in the discussion of writing models, testing writing ability is not only about language knowledge. Douglas (2000, pp. 35-40) highlights a difference between language knowledge and strategic competence. The former is further divided into (a) grammatical knowledge, (b) textual knowledge, (c) functional knowledge, and (d) sociolinguistic knowledge. These components comprise the language learner's basic knowledge about how language elements are built up, how to build a text from them, how to reach a communicative purpose with the message, and how to apply the message to different social contexts. Strategic competence, on the other hand, is not about language ability. The elements of this competence are (a) assessment, (b) goal setting, (c) planning, and (d) control of execution. These are the competences that connect language ability and the external context, and their presence is to differentiate novice and experienced writers. The construct of different writing assessments is built up of the above elements based on considerations about the assessment purpose.

Trivial it may sound, but it is clear that writing ability can be tested through writing. For this reason, traditionally, writing tests were regarded as *performance tests* (McNamara, 1996) within the framework of communicative language testing. Performance refers to the idea that test takers have to produce a performance that is observable. This view gives the basis of the differentiation between direct and indirect testing. Direct tests directly measure the abilities based on performance whereas indirect tests measure the underlying abilities defined in the construct (Fulcher, 2010). This idea hinges on a reliance on objective observation and the importance of natural language use in language tests. It has been widely thought that, if speaking skills are measured with speaking, and writing ability with writing, the test form is direct. The case of writing assessment, however, is not

that simple. As Bukta (2013, p. 39) notes, “in a performance test candidates’ ability is elicited with an instrument to arrive at a performance which raters assess using a rating scale.” Although it is possible to observe the written product, there are several underlying abilities that can only be indirectly assessed. For this reason, Bárdos (2002, p. 62) claims that all pedagogical testing is by definition indirect. There is an interaction between the participants in the rating process: raters interpret the scales, on the basis of which they assess the performance of the test taker, whose performance is determined by both the instrument and the external factors (McNamara 1996, p. 9.) (The rating procedure is detailed in Section 3.4.)

Dörnyei (1988) reviews the different trends in language testing, and he highlights the use of discrete-point tasks and integrative tasks. As a result of a strong psychometric influence, especially in the US, discrete-point testing became prevalent in the 1960s. The most well-known example of this approach is the TOEFL test (Read, 2015). These tests focus on one form, such as the grammatically correct use of a particular tense, word order, etc. Discrete-point tasks test particular elements of the language, are meticulously constructed and highly reliable. The problem with discrete-point tests is that they test isolated elements of language without context (e.g. multiple choice or matching). With the sociolinguistic turn in the 1970s, however, attention turned towards real-life based tasks and they lead to the appearance of integrative tests, which aimed to test language in context. Integrative testing, rather than focusing a single small piece of linguistic information, is trying to establish whether the test taker is capable of processing several types of information simultaneously. A typical integrative testing method is the *cloze test*, a text completion task, where several linguistic skills need to be mobilized at the same time when completing a text. In the case of testing language in context, a further distinction is made between independent and integrated task types. Although integrated tasks exist in real life especially in the context of academic writing, there are also shortcomings of this task type. The advantages and disadvantages of integrated tasks are discussed in detail in Chapter 6.

In the context of large scale, high-stakes assessment, writing is usually tested with two different task types. Although ideally a test needs to include tasks that represent different areas of language use, and they should test all the underlying abilities defined in the construct, there are practical considerations in terms of which assessment boards limit the number of tasks. Using a single writing task, however, is not satisfactory, as test takers’

abilities need to be tested “in different categories of genre, rhetorical task, pattern of exposition and cognitive demands” (Weigle, 2002, p. 65).

3.2 Second language writing and its assessment in the CEFR

The *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR) (Council of Europe, 2001) and the new *CEFR Companion Volume with New Descriptors* (CV) (Council of Europe, 2018) have been widely used as a key reference document for language testers. In addition to being a manual for assessment professionals, the CEFR is one of the best known and most used policy instruments in the European Union as regards language competence. The CEFR is a synthesis of second and foreign language learning and teaching, and assessment and thus it might serve as a basis for language test development projects (Alderson & Huhta, 2005; North, 2004), it can be used for describing different language levels, referred to in content specification and standard setting (Morrow, 2004). In this chapter, I discuss the CEFR in general, and go through the aims of its nine chapters. The dissertation provides a detailed analysis of the illustrative scales which are relevant to the writing skills in Chapter 8.

In general, the CEFR as a document of standardization aims at (a) promoting and facilitating cooperation among different countries beyond Europe; (b) provide a basis for recognition of foreign language qualifications; and (c) assists teachers, course designers and examining bodies (Council of Europe, 2001, pp. 5-6). The first three introductory chapters of the policy document define the objectives of the framework and introduce the common reference scales and the self-assessment grids. It is important to highlight that the communicative language model of the CEFR breaks away from the Four Skills Model of language knowledge introduced by Lado, (1961) (reading, writing, listening and speaking) and uses a model of communicative language activities, such as reception, production, interaction and mediation. North (2014) compares the two models and points out that the fours skills of the Lado-model are split to reception and production in the CEFR, which may be found both in written and spoken language. These receptive and productive activities are supported by phonology/orthography, lexis and grammar.

Chapter 4 of the CEFR defines the categories of language use and the language user and is recommended to be consulted in the process of designing task specifications, whereas Chapter 5 identifies the communicative language competences of the language user and is useful for test item development. There are illustrative scales provided in the

two chapters as well. The two chapters provide 50 illustrative scales altogether presented under the main categories of language proficiency which are (a) communicative language activities, (b) language strategies, and (c) communicative language competencies. The distribution of the scales is shown in Table 1.

Table 1
The Number of CEFR Illustrative Scales of Language Proficiency

	Communicative Language Activities		**Communicative Language Strategies**		**Communicative Language Competencies**
Production	Speaking	5 scales	3 scales	**Linguistic**	6 scales
	Writing	3 scales			
Reception	Listening	5 scales	1 scale	**Pragmatic**	6 scales
	Reading	5 scales			
Interaction	Spoken	9 scales	3 scales	**Sociolinguistic**	1 scale
	Written	3 scales			

As we can see in Table 1, writing performance illustrative scales appear both within production and interaction in the CEFR. These scales describe the different strategies employed in writing (planning, compensating, monitoring) and provide examples of writing activities.

The CEFR scale model and the hierarchy of the scales for the different tasks and criteria are summarised by Harsch and Banerjee (2016). They present a hierarchical model of the different language activities and scales. Production and Interaction appear as a component of both Communicative Language Activities and Communicative Strategies, which are all part of a person's Communicative Competence. Language Activities are further divided according to the different tasks learners perform, and can also be assigned different criteria on the basis of which learners are assessed. The hierarchy of scales is designed for spoken and written productive and interactive tasks. The three criteria used for assessment are Linguistic, Sociolinguistic and Pragmatic.

On the horizontal scale of the writing grid, there are topics categorised into: personal and daily life, social, academic and professional domains. Tests at the A levels

typically involve topics from personal and daily life, the B level ones select from the social domain, whereas the ones at C level use topics from the academic or professional domains. The descriptors of the vertical scales include (a) general linguistic range, (b) vocabulary range, (c) vocabulary control, (d) grammatical accuracy, (e) orthographic control, (f) sociolinguistic appropriateness, (g) flexibility, (h) thematic development, (i) coherence and cohesion, and (j) propositional precision. The CEFR supports the idea of communicative language use and provides descriptors for six proficiency levels, from A1 to C2, where the A levels are referred to as 'basic user' stage, the B levels 'independent user' stage, and the C levels 'proficient user' stage.

Chapter 6 and 7 in the CEFR are closely related. Chapter 6 is one of the two theoretically grounded chapters of the document. It discusses how a new language is acquired or learnt. Chapter 7 is a more practical one, discussing the role of tasks in language learning. Chapter 8, the other theoretical part, provides a description of the principles of curriculum design with special attention to the language learner's plurilingual and pluricultural competences "in order to deal with the communicative challenges posed by living in a multilingual and multicultural Europe" (Council of Europe, 2001, p. vii). Chapter 9, entitled *Assessment* deals with the assessment of all four skills and outlines three main issues of language testing: (a) what is assessed; (b) how performance is interpreted; and (c) how comparisons can be made.

3.3 Critique of the CEFR

Although the CEFR is used for educational and assessment purposes inside and outside the European Union, and it is the most important language policy document of the Council of Europe (Byram & Parmenter, 2012; Spolsky et al., 2014), it has also been criticised for a number of reasons. The critical reviews point out its lack of theoretical foundation (Fulcher, 2010; Fulcher, 2016; Weir, 2005b). The scales only contain 'Can Do' statements, but these statements are not rooted in any second language acquisition theory (Fulcher, 2010). As a result, the CEFR is merely a taxonomy of performance scales based on communicative language competence (Papageorgiou, 2009). This undertheorized approach has not yet been resolved with the publishing of the *Common European Framework of Reference for Languages Companion Volume with New Descriptors* (CV) (Council of Europe, 2018). The changes implemented in the CV range from the stylistic correction of

wording to rewriting or completing the illustrative scales, however, the major point of criticism regarding under theorisation remained unattended (Lukácsi, 2019b, pp. 47-52).

As for the assessment of writing, Weir (2005b) points out that the 'Can Do' statements enhance task variability but hinder setting the level of difficulty. The ambitious claim to 'common' reference scales only entails the common understanding of the teachers who took part in the development process. Therefore, "what is being scaled is not necessarily learner proficiency but teacher/rater's perception of that proficiency" (North, 2000, p. 573). Regarding scoring writing assessments, Harsch and Rupp (2011, p. 11) highlight that the writing model of the CEFR does not take into consideration the task environment and the social context of the writing process. As regards test development and validation purposes, Weir's (2005a) socio-cognitive framework, Bachman and Palmer's (2010) assessment use argument, and Kane's (2013) interpretation/use argument emphasise the importance of the context outside the test and the need for an integrated framework. It is important to stress that validity shall never be subordinated to reliability, and local contexts, procedures and stakeholders need to be taken into account (Hamp-Lyons, 2003, 163-164). Consequently, the question may arise: Why is the CEFR used in test development and standard setting? Fulcher (2010, p. 213) suggests that, although the scales lack theory and some of the descriptors are inadequate, the basic ideas can be adopted and used to build our own framework during a test development process. That is exactly what I am going to do in the second part of my thesis.

3.4 The rating procedure of writing tasks

Evaluating writing performance is a complex linguistic and psychometric task. As Eckes et al. (2016) summarise, writing assessment today "has brought together lines of development in research and theorizing on the nature of writing ability on the one hand and advances in measurement methodology and statistical analysis on the other" (p. 148). The factors that determine the assessment of the writing performance in the context of language testing are: (a) regional traditions; (b) professional background and training (McNamara & Knoch, 2012, pp. 556-558). McNamara and Knoch distinguish two traditions in writing assessment: the British and the American. They note that the UK exam boards strongly rely on the relationship of testing and teaching, especially communicative language teaching. Whereas the US tradition takes psychometric considerations into account, that is,

test formats pay more attention to psychometric characteristics and scoring validity than authentic language use.

Weir (2005a) stresses a strong relationship between context validity and scoring validity. Scoring itself means assigning a mark to a (written) performance based on a mark scheme. It is of particular importance that the mark scheme should assess the construct of the test, the raters interpret the mark schemes the same way, and internal consistency or the reliability of rating be high. In other words, subjective rating tools should be designed in a way that they ensure sufficiently high reliability. He also points out the importance of bringing raters close together in terms of their scoring and their consistency.

Erdosi (2001, p. 176) points out that variability in rater behaviour leads to validity issues in the assessment of writing products. Human raters cannot behave like computers and are never able to apply a set of detailed descriptors uniformly. Instead, they behave like readers who bring their experience to the assessment. For this reason, extensive research is needed to explore the rating process and reveal the mental processes of raters.

Research into the rating process itself in the past few decades proposed different stages that can be assigned to rating based on the cognitive processes of raters. These mental processes are best researched with qualitative methods, through which the different stages and processes can be traced. Milanovic et al. (1996) used a number of qualitative methods to reveal the cognitive processes behind the rating process of 16 raters. Their model shows both the processes themselves and the elements raters focus on during the process (Figure 8).

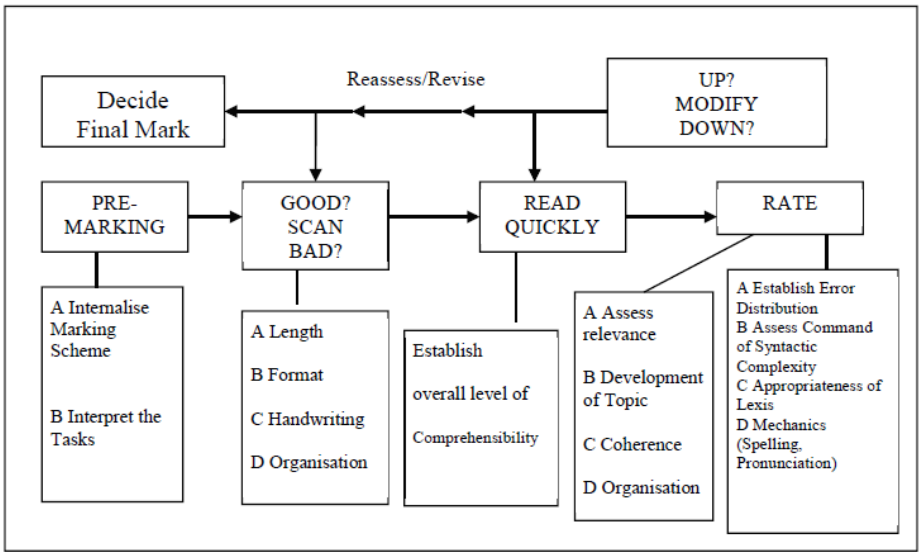


Figure 8. The four-stage model of Milanovic et al. (Milanovic et al., 1996, p. 95)

Milanovic et al. designed a four-stage model: (a) pre-marking, (b) scanning, (c) quick reading, and (d) rating. During pre-marking, the raters internalise the marking scheme and interpret the task. In the next stage, they scan the writing product and try to decide if it is good or bad based on surface features, such as length, format, handwriting and organisation. The third stage entails a more detailed reading when raters look at overall comprehensibility through which they enter the fourth, rating stage. In this phase, they assess both content (relevance, development of topic, coherence, and organisation) and linguistic features (errors, command of syntax, lexis and spelling). In addition to the stages, Milanovic et al. identified four different reading approaches raters use: (a) principled read, (b) pragmatic read, (c) read-through approach, and (d) provisional mark approach. Principled and pragmatic read both involve reading the script twice but with different purposes. The former is done with bearing the scoring criteria in mind, whereas the latter is employed when the rater identifies a difficulty while reading the script. The read-through approach is the most superficial of the four as it involves only one reading. Although the provisional mark approach is also characterised by a single reading, it is more in-depth in nature because it is focusing on the merits of the text and the efforts of the candidate.

As Bukta (2013, p. 84) notes, models of the 1990s tried to take into account how raters' background (thinking, expertise, training) influenced the scoring process. Wolfe (1997) tries to differentiate between raters based on proficiency levels and he bases his model on different thinking processes (Figure 9).

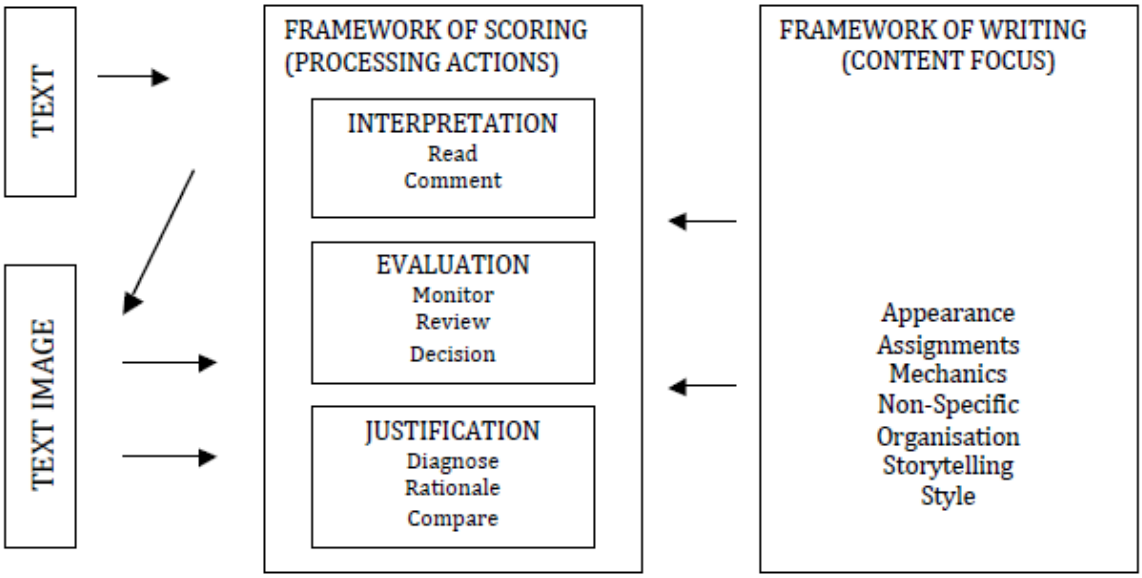


Figure 9. Model of scoring based on rater's cognition (Wolfe, 1997, p. 89)

Wolfe's model of scoring cognition suggests that raters first read the text, interpret it, and create their own image of the text. The scoring stage is affected by this text image and the content-focused writing product; in other words, scoring is a combination of two frameworks: (a) the framework of writing and (b) the framework of scoring. Wolfe was interested in comparing the different rating styles of less proficient and proficient raters. He found that less proficient raters stop more frequently in the rating process and tend to make decisions at certain points, as opposed to proficient raters, who usually assign a score to the writing product at the end of the process.

Lumley (2002) also used think-aloud protocols, but he worked only with proficient raters. He investigated whether raters agree on the meaning of the scale components, whether they follow the set of components or they go outside the scale. His model (Figure 10) is built up of three stages the raters in his study tended to follow in a similar way. The three stages are: (a) first reading, (b) scoring the categories of the scale, and (c) final consideration.

Stage		Rater's focus	Observable behaviours
1.	First reading (pre-scoring)	• Overall impression of text: global and local features	• Identify script. • Read text. • Comment on salient features.
2.	Rate all four scoring categories in turn	• Scale and text	• Articulate and justify scores. • Refer to scale descriptors. • Reread text.
3.	Consider scores given	• Scale and text	• Confirm or revise existing scores.

Figure 10. Stages in the rating sequence (Lumley, 2002, p. 255)

These three stages show similarities with an earlier model from the 1980's developed by Freedman & Calfee (1983) and with Wolfe's model of rater cognition as well in that raters assess a text image they create through reading the text, adding their experience and background knowledge. Weigle (2002, p. 71) notes that rater expectations have been proved to influence their judgements and that, in some cases, external factors may be as important as the written product itself.

3.5 Raters of subjectively marked tasks

Despite the increasing number of automated essay scoring (AES) in the context of language testing, computerised assessment receives harsh criticism (Van Moere & Downey, 2016). It seems today that human raters employed to assess writing products will not be replaced in the near future (Inoue et al., 2018; Weigle, 2002). For this reason, it is of particular importance that the subjective nature of human rating should be minimised. To be able to do this, we have to understand the cognitive processes of raters during the rating process.

Another issue to be discussed in connection with the research into the rating process is the method of data collection. As we saw it above, verbal protocols are the most widespread research technique in both theory building and rater research (Lumley, 2002; Weigle, 1994; 1998; Wolfe, 1997). As mental processes are not observable directly, the only way to learn about raters' thoughts is through introspective methods. Using this type of qualitative data collection method means using raters' verbal report of their thoughts either retrospectively or concurrently as they perform a task (Mackey & Gass, 2005, p. 77). These verbal reports are recorded and transcribed and analysed to reveal the different aspects raters pay attention to in the scoring process. An advantage of verbal protocols is that they reveal the participants' thoughts and through a think aloud process, the researcher can actually experience the process being carried out. The information is not consciously filtered, it is more likely to be about what participants actually do, rather than what they believe they do (Barkaoui, 2011, p. 52). The method, however, has its limitations. Smagorinsky (1994, pp. 3-4) points out that, in addition to verbal protocols being labour intensive for the researcher (recording, transcribing, coding and analysing data), it might be difficult for the participants to verbalise their thoughts when they are performing the task of rating. Despite the limitations of the data collection method, this is the only way employed to gain insight to the cognitive processes of raters (Methods of data collection are discussed in detail in Chapter 5).

As the scoring models revealed, rater influence or rater variability is part of writing assessment. It is often referred to as *construct irrelevant variance*, in other words, a factor that is not relevant to the construct being measured (Messick, 1994, pp. 14-15). The most common rater effects are (a) rater severity/leniency, (b) central tendency, (c) halo, and (d) rater bias (Eckes et al., 2016, p. 155). Severity/leniency means that the scores given by raters are either too high or too low compared to each other, or to the standardised mark.

Raters show central tendency if they repeatedly give marks around the middle and do not differentiate between strong and weak test takers. The halo effect refers to raters assigning scores to a text based on one single criterion, and rater bias is the case when raters are externally influenced and thus become inconsistent regarding their severity/leniency. Consistency is key in large-scale assessments, which means that leniency or severity on their own are not regarded as problematic as they can be compensated for with statistical methods (Bukta, 2013, p. 91). However, in order to reach consistency, repeated rater training is essential. Lumley (2002) argues that "rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential" (p. 267). Rater training must be systematic in order to reach consistency and high rater agreement. The training provided is essentially monitored and raters should always be provided feedback (Weir, 2005a).

Ratings may be subject to personal judgements, even "trained experienced raters have been shown to differ systematically in their interpretation of routinely-used scoring criteria" (Eckes, 2009, p. 5). Nevertheless, research into rater training is an important issue in L2 writing assessment. Research in the field usually focuses on differences between experienced and inexperienced raters, and how consistency and reliability can be enhanced through training. Weigle (1994) used verbal protocols in her study with inexperienced raters. She looked at their rating style and how they understood rating criteria before and after training. Her analysis revealed that inexperienced raters highly benefitted from training. Most importantly, the training provided them with the possibility of comparing themselves to a reference group. Weigle (1998) compared the severity of experienced and novice raters and found that novice raters tended to be more severe, but at the same time much more inconsistent in their decisions. Her results are similar to those of her earlier research: after training, the magnitude of difference between experienced and novice raters was smaller. We may conclude that rater training improves reliability, but it does not eliminate individual differences regarding severity/leniency.

Eckes (2008) studied the trained, experienced raters at the TestDaf Institute and found that even highly trained raters show different rating styles. Based on this observation, he described six different rater types: (a) the syntax type, (b) the correctness type, (c) the structure type, (d) the fluency type, (e) the non-fluency type, and (f) the non-argumentation type. The different rater types regarded the different rating criteria with unequal attention. He warned that "high interrater reliability could simply be due to these

raters' type-specific points of view regarding the weight of scoring criteria" (Eckes, 2008, p. 179). We may conclude that rater training is an important factor in increasing rater reliability, yet rater variability cannot be eliminated, only better explained.

In rater training sessions, it is important to provide raters with feedback. First and foremost, they should be communicated the amount of variability that is acceptable; in addition to that, raters should be given feedback on how they perform compared to the rest of the group (Weigle, 2002; Weir, 2005a). Similarly to rater training, rater feedback has mixed results. Elder et al., (2005) Knoch (2011), and Wiggelsworth (1993) report that providing feedback does not guarantee that the raters are interpreting the scale in the same way. Although raters are generally positive about the feedback they get, it does not usually have a long-term effect. For this reason, eliminating the differences entirely between raters is not possible. Even if it were possible, elimination is not desirable; it is more advisable to use feedback to enhance self-consistency of raters (McNamara, 1996).

Another way of dealing with rater variability issues and increase reliability is using double blind rating, i.e. having two raters who independently assess the same performance. In this case, rater reliability can be demonstrated through a small difference between the two independent raters. Apart from self-consistency (intra-rater reliability), there needs to be consistency between the two raters as well (inter-rater reliability) (Weir, 2005a). As Eckes et al. (2016) suggest that both consistency and consensus should be observed when computing inter-rater reliability (pp. 155-156). The consistency index shows whether the relative ordering of the performances is similar. A high correlation between the ranking indicates that they rank the test takers more or less the same. However, high correlation does not show whether the two raters understand and use the scale in the same way. To be able to demonstrate this, we have to observe the consensus index, in other words, the degree (percentage) of exact agreement.

3.6 Rating scales and checklists

The assessment procedure of writing tasks has always generated interest in linguistic research. The tool most assessment related handbooks describe for the assessment of writing products is the rating scale (Alderson et al., 1995; Bachman & Palmer, 1996; McNamara, 1996; Shaw & Weir, 2007; Weigle, 2002, Weir, 2005a). At the same time, the fallacy of subjective marking of learners' and test takers' writing performance and the need for more objective, i.e. more consistent assessment has been repeatedly raised by a number

of publications (Eckes, 2009; Knoch, 2009; Knoch, 2011, Lukácsi, 2017; 2018; McNamara, 2000). The CEFR document also states that "in 'rating on a scale' the emphasis is on placing the person rated on a series of bands". These bands are displayed vertically and the raters place the performance of the test taker on this vertical scale. As opposed to this, the binary choice decisions of a checklist cover different aspects of language production, making it suitable to decide whether a test taker is at a certain level. In other words, rating with a checklist is "showing that relevant ground has been covered" (Council of Europe, 2001, p. 189). The emphasis for the rating scale is vertical while this emphasis for a checklist is horizontal, which makes the latter more suitable for level testing. In this section, I discuss the use of rating scales in assessing writing and look at checklist-based assessment as a possible alternative.

In order to assess what the writing construct defines, raters need to refer to a detailed set of descriptors. The descriptors of the locally developed rating tools specify the details of what performance is expected from test takers and are most often placed on a scale (Van Moere, 2014). Research into the rating process, rater training, and rater agreement revealed that it is difficult to bring raters close to each other in their judgements, so it is essential to provide raters with a well-designed rating scale (Weir, 2005a). Scales have multiple functions in language testing: they are used to assess the test taker's performance, they serve as a guidance for raters, and provide test designers with information on test specifications and the construct (Bukta, 2013, pp. 52-53). The three scoring types that appear on scales are: (a) holistic, (b) analytic, and (b) trait based (Hyland, 2004, p. 226).

Holistic scoring means that raters assign one score to a performance which is based on a single impression (Weigle, 2002). The advantage of holistic scoring is that developing the scale and training raters are less time consuming. Moreover, rating with a holistic scale focuses the reader's attention, which makes the reading process more authentic (Hamp-Lyons, 1995; White, 1984). However, it is important to point out that Hamp-Lyons specifically writes about assessing L1 writing, which may be significantly different from testing L2 writing performance. Weir (2005a, p. 183) highlights that the difficulty in using holistic scales for L2 writing performance is that test takers may show different levels in different criteria. In his criticism, he refers to Bachman (1988) who contends that holistic writing scales do not pay attention to the acquisition order of various language elements.

Analytic scoring is explicit about detailed, individual aspects of written performance. It involves a number of different criteria on a rating scale, including range of grammar and vocabulary, content, coherence and cohesion, and structure. The detailed analytic scale should be designed based on evidence and should be appropriate to the discourse community and it must suit the purposes of the test (Weigle, 2002, pp. 114-120; Weir, 2005a, pp. 183-188). The analytic approach, provided the scoring tool is appropriate and explicit, may give a more detailed evaluation of the test taker's performance. At the same time, it may enhance the reliability of inexperienced raters (Weir, 1990). The detailed nature of analytic scales might be an advantage but some of the disadvantages also lie in this feature: they are difficult and time-consuming to develop, at the same time it might be difficult for raters to pay attention to all the criteria (Hughes, 2003, pp. 103-104).

Trait-based scoring can be divided into two categories: primary trait and multi-trait. Primary trait scales are designed for a specific task; they are highly detailed using several categories to help raters. Being time-consuming, primary trait scoring is not used in large-scale language testing (Eckes et al., 2016, p. 154; Weigle, 2002, pp. 110-112). Multi-trait scoring, on the other hand, combines the advantages of trait-based and analytic scoring (Hyland, 2004, pp. 229-232), yet it is highly time-consuming.

In addition to the widely used rating scales, language testing researchers suggest a number of alternatives to eliminate the disadvantages and the inherent fallacies of rating scales (Lukácsi, 2018; 2020). To improve the validity of rating, Fulcher et al. (2011) developed a Performance Decision Tree in response to the shortcomings of rating scales. This tool offers binary choice decisions, which makes it easy to describe and evaluate a communicative task. The individual yes/no decisions are independent of each other and cover different aspects of language production. Similarly, in order to increase objectivity, Struthers et al. (2013) designed a checklist for evaluating and assessing cohesion in children's writing, and Kim (2011) developed a 35-item checklist to assess academic English. Common traits of the performance decision tree and the checklist are: (a) finely tuned and level-specific items, and (b) binary-choice decisions. These features help raters find concrete evidence to assess language and writing quality (Brindley, 2000). Finding concrete evidence is essential in assessing writing products. Mickan (2003) when highlighting the issue of inconsistency, points out the lack of evidence in the IELTS examiners' writing task ratings: he found that the different levels of performance assigned to writing products did not show distinguishable features in terms of lexico-grammatical

performance. The reason for this is that despite the use of analytic scales, raters looked at texts as a whole rather than paying attention to individual components. The advantage of the checklist lies in the binary decisions, which lead to a more reliable and objective assessment. Furthermore, the checklist items help maintain the construct relevant nature of the assessment.

The aim of the dissertation is to establish the validity of the writing tasks of the locally developed EAP test, with a special focus on the scoring validity of the discussion essay within the academic domain. This aim fits into the parallel research project of the exam centre that focused on solving the problems of rater inconsistencies that appear in the assessment of the subjectively marked tasks and aimed at increasing the reliability and fairness of test scores. Earlier research proposed a checklist-based rating tool (Lukácsi, 2017; 2018; 2020) for the assessment of the B2 level written and spoken products. Chapter 8 of the dissertation presents the comparison of rating on a scale and rating on a checklist in the context of the Euroexam EAP test development project, and aims to design a level and genre specific checklist for the assessment of the C1 level discussion essay within the academic domain.

Chapter 4: Test Development

This chapter discusses the fundamentals of test design and development for large-scale assessment. The first part of the chapter presents the procedure of task development, the different phases and considerations of the stakeholders within the Assessment Use Argument (AUA) framework (Bachman & Palmer, 2010). The second part discusses concepts of validation in language testing. After examining general and specific concepts of validity, the particular frameworks covered are the characteristics of test usefulness (Bachman & Palmer, 1996) and the socio-cognitive validation framework (Shaw & Weir, 2007; Weir, 2005a).

4.1 The development process of tasks for large-scale language tests

Language test development and task design require a lot of time and effort and meticulous design. Alderson and Clapham (1995, p. 185) claim that there is no significant difference between designing a high-stakes, large-scale assessment and a classroom test. Bachman and Palmer (2010) share this view and emphasize the need for theory based test development in low stakes test design, however they argue that the test development process does depend on the scale of the project. The procedure for developing a large-scale test involves more people (test development experts, judges, students to be tested) and the decisions to be taken are much more important and formal.

The procedure for language test development is described in a number of handbooks (Alderson et al., 1995, Bachman & Palmer, 2010; Read, 2015), and they all emphasize that test development can be divided into different stages. The main stages that are usually differentiated are (a) design, (b) operationalisation, and (c) administration (Weigle, 2002, pp. 77-78). Bachman and Palmer (2010), though, propose five stages for test development, which are (a) initial planning, (b) design, (c) operationalisation, (d) trialling and (d) assessment. In my research, I draw on their five-stage model.

The initial planning stage includes the considerations of a suitable measurement tool and the review of the ensuing financial and human resources. The next stage is the design when a general plan is designed to serve as a guide for the development team. This is followed by operationalisation, when the test blueprint is drawn up together with proposed test tasks and items. Then, the test tasks are trialled with test takers who are similar to the group of people the test is designed for. Trialling (as used by Bachman and Palmer, 2010) can involve piloting and pre-testing. These procedures are used in test

development to see how examination materials work in practice (Weir 2005, p. 206). The trialling stage usually informs the previous stage, in other words, the stages follow each other in an iterative manner: operationalisation and trialling may go on until the test material works as expected. The last stage follows only if the test material is of acceptable quality; however, constant monitoring through statistical analysis is essential. The operationalisation of the measurement tool is an important process in the context of localisation, which will be discussed below.

The initial stage of test development, according to Bachman and Palmer (2010), involves a number of decisions that affect the course of the entire test development project. Major policy decisions are to be made for high-stakes language tests at this stage, for this reason, their importance is high for the different stakeholders (Menken, 2017). The first professional decisions are usually about producing the *mandate* (Lynch & Davidson, 1994, p. 728). The mandate is usually a policy document based on the needs of stakeholders, decision makers, or administrative bodies. The mandate defines the level, the purpose and the general framework of the test. It is important to highlight that the developers' considerations that appear in the mandate are valid within its context.

After the initial considerations and formulations of policies, resources, and the target population in the mandate, in the design stage the construct itself needs to be defined. Fulcher (2010, p. 96) conceptualises the construct with the help of the metaphor of 'design patterns in architecture'. The construct measured by a language test is made up of the qualities, abilities and skills considered to be ideal for a specific purpose (Vígh, 2005, p. 328). Construct definitions can be curriculum-based or theory-based (Weigle, 2002, p. 79). Curriculum-based constructs are based on a specific course syllabus, whereas theory-based constructs are based on a theoretical model of the skills to be tested. High-stakes proficiency tests, which are in the focus of my research, fall in the latter category. Theory-based constructs are not easy to define because the abilities that underlie the test takers' performance cannot be directly observed (Fulcher & Davidson, 2007, p. 36). In addition, it is very difficult to re-create real-life circumstances under exam conditions therefore construct irrelevant features need to be defined very carefully.

Constructs are outlined in terms of several models or frameworks and include taxonomies about language ability (Bachman & Palmer, 1996; Council of Europe, 2001; 2018; Weir, 2005b). Test makers will formulate their design statements accordingly. This is also the moment when test designers must collect information and evidence about the

targeted test takers and their background. In case of Languages for Specific Purposes (LSP) tests, such as English for Academic Purposes (EAP) in my research, defining background or topical knowledge is essential and should be integrated in the construct definition (Douglas, 2000). The last phase of the design stage is the consideration of the six aspects of test usefulness: (a) reliability, (b) construct validity, (c) authenticity, (d) interactiveness, (e) impact and (f) practicality (Bachman & Palmer, 1996). The consideration of these characteristics is an essential part of validation research and will be discussed in detail in Section 4.2.

The third stage is called operationalisation and it is the moment when the initial considerations and the design statement are turned into actual test specifications, test tasks and items. Creating detailed test specifications is the core of the test development process, as they are going to be used in the future for creating tests for testing the same construct. They also have practical consequences as they will be used to check future test batteries against them (Bachman & Palmer, 2010), and to create an item bank (Davidson & Lynch, 2002). Test specifications, however, are not only for the development team or language testing professionals. Alderson et al. (1995) suggests creating different specifications for different audiences, such as teachers, test takers, item writers, and policy makers. Douglas (2000, p. 249) lists the essential elements of specifications for wider audiences as follows: (a) a description of the content – including the number of tasks, (b) the time allotment, (c) the rating criteria, and (d) sample tasks.

The proposed test tasks need to be tested. For this purpose, it is essential to have participants in the trial stage who are as similar to the target population of the test as possible. In order to make amendments, it is also recommended to get information based on the suggestions of the different stakeholders (schools, teachers, parents, etc.). It is important to stress that after the trialling phase, both the test tasks and the specifications are to be revised. Although Lynch and Davidson (1994, p. 731) describe test specifications as a dynamic process rather than fixed, revision and improvement should always be well-grounded and well-documented. The ALTE *Manual for Language Test Development and Examining* (ALTE, 2011), which complements the CEFR (Council of Europe, 2001), and was designed to support test development processes, highlights that both the rating scales and the mark schemes are to be tried out, and the results should be analysed. In order to achieve high reliability and comparability, scales and mark schemes must not be altered after the trials.

There are different methods for try-outs. In the literature, there is a distinction drawn between piloting or trialling and pre-testing. Butler et al. (1996) use the two terms with inverse meanings. In my research, I draw on the literature according to which piloting or trialling is a small sample trial using strategies of qualitative inquiry, while pretesting is carried out in a testing environment using a sample which is similar to the target population (Alderson et al., 1995; ALTE, 2011; Council of Europe, 2009; Fulcher & Davidson, 2007; Read 2015; Weir, 2005a). In the dissertation I avoid the use of 'piloting' and will use the terms 'trialling' and 'pretesting' in order to avoid any confusion with 'pilot study' in Stage 4 of the research.

Trialling of test materials is necessary (a) to eliminate ambiguities in the test, (b) to check the clarity of the questions and their instructions, and (c) to estimate the difficulty level of the task based on the test taker's comments and the time load (Council of Europe, 2009, p. 91). In the course of trialling, both qualitative and quantitative information can be gathered from stakeholders, however, small-scale trials, such as using cognitive labs with 3-6 participants also suffice (Paulsen & Levine, 1999; Zucker et al., 2004). The cognitive interviews can be used to gain insight to test takers' mental processes and get invaluable feedback on the quality of the test. Pretesting on the other hand, is used to get information on how the test material functions as part of a planned examination with a population that is representative of the target population (Council of Europe, 2009, p. 92). Apart from gathering item-level quantitative data, pretesting can also be used to determine the internal validity of the test, i.e. how linked items work together. The methods I applied in try-outs are described in detail in Chapter 5.

The fifth, last stage of the development process entails putting the test tasks to use. Administering tests is highly bureaucratic in nature, however, the developer's instructions are to be followed so that the test-takers of all exam sessions take the same standardised test. Administration might be a burden for the stakeholders, but it is in the interest of all parties that delivering the test goes smoothly. Checking the number of copies, their delivery, providing qualified invigilators, making sure that the examination rooms are suitable for the test are tiresome and monotonous tasks, but they all serve the greater purpose of providing a fair test with comparable results. The strict rules are set in order to "protect score meaning, validity, and the fairness of the outcome of the test and any decisions that might be associated with the results" (Fulcher, 2010, p. 254). It is important

to control the construct irrelevant factors to maximise test taker performance and make circumstances for data collection optimal.

4.2 Test usefulness

Ideally, language tests resemble authentic language use and provide useful information about the test taker's language abilities. The different stages of the test development process described above all observe the question of usefulness, in other words whether the test measures the construct defined earlier, and whether the scores reflect the knowledge, skills and abilities of the test taker (Fulcher, 2010).

Bachman and Palmer contend that "the most important quality of a test is its usefulness" (1996, p. 17). They (Bachman & Palmer, 2010) introduced later the principle of Assessment Use Argument (AUA) with six characteristics to consider in test development. They argue that, for producing a valid test, one must secure the (a) reliability, (b) the construct validity, (c) the authenticity, (d) the interactiveness, (e) the impact and (f) the practicality of the test tasks. These criteria mean, respectively, that test makers have to consider (a) whether the results would be the same if the test is taken a number of times; (b) whether they do measure what they are supposed to measure; (c) whether the test tasks are similar to real life tasks. In addition to these, we also have to consider (d) if there are other skills which might be interacting when taking the test, and (e) how the test affects the different stakeholders. Last but not least, an equally important aspect of the test is (f) how easy it is to implement. In this section, I introduce the qualities of test usefulness for language testing, especially the assessment of writing, as proposed by Bachman & Palmer (1996; 2010). Since validation is a central topic of the dissertation, the different concepts of validity are detailed in Section 4.3.

Reliability means the consistency of measurement. It was Lado (1961) who provided this classic definition of reliability. A test is considered reliable if test takers get the same results if they take the test twice, or even if they are scored by different raters. There are different methods to calculate reliability, which are detailed in Chapter 5. In the course of a test development project, it has to be assured that the scores are generalizable. The score on a test has to be the adequate reflection of the test taker's abilities and the results should be consistent across different assessment periods (Knoch & Elder, 2013). If reliability is high, the same test taker should achieve the same result under the same conditions. Absolute identity, however, is excluded as neither the absolute identity of the

circumstances, nor the exact same performance can be assured between different test administrations. Thus, reliability means the reliability of the instrument itself, which operates in a population that retains roughly its original characteristics but in different circumstances. The true score of a test is always the combination of the test taker's actual score and the measurement error (SEM) (Bárdos, 2002, p. 38). As for the assessment of writing, a true score is a slippery concept because it is affected by different factors. In case of writing assessment rated by humans, rater behaviour can be a source of variance (Weigle, 2002). In the case of high-stakes writing assignments, for securing reliability, the following should be observed without sacrificing economy: (a) controlled essay reading, (b) using a scoring critera guide, (c) assessing anchor papers as examples, (d) checking the reading process, (e) utilising multiple independent scoring and eliminating the discrepant scores, and (f) evaluation and record keeping (White, 1984, pp. 404-405).

Validity within the classic model refers to the appropriateness of the test: it shows whether the instrument really measures what it claims to be measuring. Validity as a concept appeared as early as the 1950s in language testing (Cronbach & Meehle, 1955). They stressed the connection between reliability and validity and focused on how test scores reflect the abilities of the test takers. However, according to recent developments (Bachman & Palmer, 1996), it is construct validity that is used to describe "the degree to which empirical evidence and theoretical rationales support the *adequacy* and appropriateness of *inferences* and *actions* based on test scores" (Messick, 1989, p. 13). As Cronbach and Meehle (1955) point out, one does not validate a test "but an interpretation of data arising from a specific procedure" (p. 447). For this reason, it is important to define the abilities and also the domain of the particular language use in the construct. The process is referred to as construct validation and is divided to five steps. The steps were summarised by Kane (2019) as follows: (a) specifying the intended interpretation and use of the test, (b) designing an assessment that fits the intended interpretation and use, (c) identifying challenges, and making revisions if needed, (d) examining the design for sources of bias or irrelevant variance, and making further revisions if needed, and lastly (e) developing an argument leading from scores to the interpretation/use.

Authenticity, as defined in Bachman and Palmer's (2010) Assessment Use Argument (AUA) refers to target language use (TLU), which means that authentic language tests assess how the language is used in real-life situations. It might seem that authenticity in AUA offers a simplified understanding, but Bachman (1990) also claims

that authenticity is "a function of the interaction between the test taker and the test task" (p. 117), which leads us to the abilities of the test taker that are defined in the construct. Bachman (1991, pp. 690-991) defines this criterion as interactional authenticity, and he introduces the idea of situational authenticity, which is defined by the relationship between the test task and the real tasks of TLU. The notion of authenticity as used by Bachman and Palmer (2010) is not a simplified one, but it breaks away from the dichotomous idea of authentic (taken from real life) and inauthentic (written for testing purposes) texts, but rather focuses on how the test taker is using the language or going to interact with the language in future real life tasks.

Interactiveness is a characteristic of a test that is related to authenticity. Bachman (1991) and Milanovic (2002) speak about interactional authenticity, and Bukta (2014, p. 47) also discusses interactiveness as a type of authenticity. In addition, interactiveness is also connected to construct validity in so far as it is about the test taker's characteristics and abilities as defined in the construct, and their interaction with the test task. The characteristics of the test taker are important because, as it is discussed above, language ability that is tested by language tests is not only about the linguistic code but rather about how this code is used in real life. Thus, the characteristics of the test taker to be tested include strategic competence, topic knowledge and affective schemata (Weigle, 2002, p. 53).

A highly interactive writing task involves goal setting, planning and assessment of various facets, such as self-assessment during task completion. An interactive task should also be interesting to the test taker. Apart from topic knowledge, emotional engagement is especially important so that the high affective filter should not hinder producing language (Krashen, 1985).

Impact is the effect a particular test has on the individual (micro level) or as a social group (macro level) (Bachman & Palmer, 1996). The impact of a test needs to be borne in mind in the planning phase, as such it is closely connected to the construct. Fulcher and Davidson (2007, p. 51) suggest that development projects should be effect-driven. This entails a twofold process. As impact works both backwards and forwards, it has an effect on the stakeholders of a test and the actual design as well.

Impact, in the sense of 'test effect', is also referred to as 'washback' in the literature (Weigle, 2002, p. 54). It also works at the two levels. At the macro level of society,

decisions based on test scores have washback effect on education systems, whereas at the micro level of the individual, washback affects learners and teachers. The washback effect can be positive or negative. Traditionally, a teacher starting to use practices that enhance language ability and at the same time promote test preparation, washback is considered positive. As opposed to this, a typical negative washback effect is when teachers are only "teaching to the test" (Fulcher, 2010, p. 6). Although the notion of washback is widely used in the context of test development and language testing, washback intentions of test designers are difficult to study. Education environments are very complex, so it is not easy to find the reasons for a change in classroom practices. It requires very careful triangulated design to be able to trace the data that would provide evidence for intended washback (Wall, 2005; Wall & Horák, 2006). I consider a positive washback effect as a potential implication of the use of checklist-based assessment at Euroexam International, which could be a topic of future research.

As Weir (2005a, p. 38) points out, research into impact and washback clearly help ensuring ethical language testing and fairness; moreover, help testers to meet the demand of critical language testing view (Shohamy, 1993; 2001). She argues for the need to "develop critical strategies to examine the uses and consequences of tests, to monitor their power, minimize their detrimental force, reveal the misuses, and empower the test takers" (Shoshamy, 2001, p. 131). Her political approach demands "democratic perspectives of testing" based on equal opportunities (pp. 129-158). Weir (2005a) suggests that validity frameworks should be observed in detail in order to enhance test fairness.

Practicality is the last element to consider in the Assessment Use Argument (AUA) model. It is defined as the relationship between the resources that are required for the development and administration of the test and the resources that are available for these activities. A test is practical if the available resources are greater than the resources needed for implementation. Bachman and Palmer (1996, p. 36) divide resources into three categories: (a) material resources (room, paper, copier, etc.), (b) human resources (administrators, invigilators, interlocutors, etc.), and (c) time (allocated time for test, time needed for rating, etc.). It is important to consult the relevant stakeholders regularly on the aspects of practicality to make sure that the examining process is reasonable compared to available resources.

As for the development phase, practicality also has to be considered in connection with test validity. A longer test may increase validity and reliability, however,

administering it might also increase the chances of errors in the measurement. There are inconsistencies in test design that testing experts have to put up with, as it is impossible to eliminate them completely. Practicality, therefore, may come first in certain contexts, which does not mean that it overwrites professional requirements, but there should be a healthy balance between the six characteristics of test usefulness.

4.3 Validity in language testing

As validity is a central concept of the dissertation, I provide a more detailed discussion of validity in the context of developing writing tasks for high-stakes tests. Reliability and validity in the early days of language testing were two separate concepts, often competing with each other (Weir, 2005a). The primary concern for testers back then was construct validity, i.e. to make sure that the test appropriately measures what it is supposed to measure. This meant that reliability, i.e. the consistency of measurement was not the main guiding principle of test developers. In early theories, argues Cronbach & Meehl (1955), validity was divided into three subtypes, which were considered to exist on their own. (a) criterion related validity, (b) content validity, and (c) construct validity.

Finding evidence for criterion-related validity means that test developers are interested in one criterion on the basis of which decisions can be made. (Fulcher & Davidson, 2007, p. 4) An example in connection with EAP writing skills would be investigating the question whether performance in an EAP test can predict how students are going to cope with the writing demands of their studies in the future. Thus, a further subtype of criterion-oriented validity is predictive validity. Content validity is about the domain of the test and finding out whether the test tasks represent that particular domain. Fulcher (1999, pp. 222-223) points out that it can be done through describing the test takers, finding out about their needs and sample form the target domain. Finally, establishing construct validity is about analysing and understanding what we wish to test.

It was Messick (1989) who introduced a unified view of validity. Integrating the content, criteria and most importantly, the consequences of a test, he defined validity as the evidence for the consequences of score interpretation. He argued that a test cannot be valid if it is not reliable: it is only possible to make inferences about language ability, and decisions based on test scores if it is clearly defined what the test measures. Kane (2016) also stresses the importance of evidence, as he believes validity claims “are not self-evident, and therefore they require evidence for their justification” (p. 64). Alderson et al.

(1995) also point out that a test is never valid if it is not reliable. On the other hand, a reliable score on its own does not always mean that a test is a valid measurement of language ability. A classic example for this is the use of independent, multiple-choice grammar items, which always give highly reliable scores, but the task type is rarely believed to be a valid measurement of foreign language ability. It is important to stress that Messick's view is a relative concept. He stresses that "validity is a matter of degree" (1989, p. 33) and he argued that it is based on an evaluative process.

Messick's unified view of validity takes into account both score interpretation and its social consequences. Within the unified concept, he stresses the fundamentally social nature of assessments and distinguishes different aspects of validity (Figure 11). The novelty in his approach lies in the second row of the figure, in the introduction of consequential basis, presenting new implications for language testers to consider. He argues that language tests have a consequential basis that is never value-free, thus the social consequences of its impact always have to be observed. This is an important consideration that I am going to take into account in the course of the development of the assessment tool. In the context of establishing the scoring validity of the Euroexam EAP test, both its value implications and its social consequences are relevant for the stakeholders.

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance / utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Figure 11. Aspects of validity (Messick, 1989, p. 20)

Messick's concept unites the interpretation of appropriateness, meaningfulness and score-based inferences (Messick, 1989; 1990; 1994). He defines six further aspects of construct validity: (a) content, (b) substantive, (c) structural, (d) generalizability, (e) external, and (f) consequential aspects. The content aspect of construct validity includes evidence of content relevance, representativeness and technical quality. The substantive aspect refers to the theoretical rationales along with empirical evidence. The structural aspect highlights the

fidelity of the scoring. The generalizability aspect examines the extent to which score properties and interpretations are generalizable. The external aspect includes evidence from multitrait-multimethod comparisons. The consequential aspect acknowledges the implications of score interpretations, as well as the potential consequences of test use (Messick, 1994 pp. 11-15).

Construct validity originally was presented by Cronbach & Meehl (1955) as an alternative to the criterion and content models to be used for constructs defined by a theory. In recent literature, however, it is often used as an umbrella term, sometimes interchangeably with validity (Weir, 2005a), and together with other types of validities that complement the notion.

Although the complex, unified concept of validity (Messick, 1989) is still in use, other validity types, such as face validity, content validity, criterion-related validity, consequential validity have been introduced. Hamp-Lyons (2003, pp. 164-166) refers to these as "the new validities". Face validity is a superficial quality of the test, it refers to whether test takers perceive the test as an authentic test. Content validity (or context validity in Weir, 2005a) is about finding out whether the test task is a representative of the target language use. Criterion-related validity is concerned with finding relationships with other tests or measures concerning a particular criterion. Finally, consequential validity challenges biased, unfair tests, and is concerned with the equal treatment of test takers.

Instead of defining validity in theory, a more pragmatic approach has been introduced, the argument-based approach, which argues for collecting evidence in terms of different criteria. This approach calls for collecting evidence in support of the uses and interpretations of test scores (Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 2006). As Kane has put it forward, "to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. […] Therefore, validation requires a clear statement of the proposed interpretations and uses" (2006, p. 23). In testing writing ability, this entails that we have to find evidence for construct validity in the following three ways: (a) the task must elicit the type of writing that we want to test, (b) the scoring criteria must take into account the components listed in the construct, and (c) the raters must always observe these criteria when scoring (Weigle, 2002, p. 51). In the course of generating validity evidence, both qualitative and quantitative methods should be used, which are discussed in Chapter 5.

The recently developed framework that follows the argument-based approach in validity theory and stresses collecting evidence of test validity is the socio-cognitive framework originally proposed by Weir (1988), elaborated in Weir (2005a), and extended in Shaw and Weir (2007). It has been used by the CRELLA Research Institute in a number of exam development projects (University of Bedfordshire, 2019). The framework provides an approach to validation research which brings together the social, cognitive and scoring dimensions of language use. Chan (2013, pp. 46-50) points out that Weir's work is ground-breaking in developing an evidence based socio-cognitive validation framework which combines the test takers' underlying cognitive abilities, the context of language use and the process of scoring operationalised in the language tests. Collecting evidence of individual validity components may be collected *a priori (*before the test event) and *a posteriori* (after the test event). The *a priori* stage is highlighted in the framework as this is what allows establishing connections between the theory and the actual test. He claims that *a posteriori* statistical data and score interpretations do not generate ample evidence on their own. It is clearly better grounded if the construct is defined before, which would give an adequate basis for statistical analysis. He also argues that test construct can be better defined by the cognitive processing involved in language use in real-life (Weir, 2005a, pp. 17-18).

Weir's framework is preferred by test developers and researchers as it focuses on systematic analyses of test input and output, from both linguistic and psychometric perspectives. It is unique in trying to grasp the abstract concept of the construct by distinguishing the cognitive and context elements during the test development and the testing process. It emerged in Weir (2005b) when pointing out the shortcomings of the CEFR (Council of Europe, 2001), and arguing for the need of a theory base framework in language test development (see Chapter 3). Exam boards and development teams of high-stakes tests need to demonstrate how they meet the validity requirements and how they operationalise these in their tasks at different levels of proficiency (Shaw & Weir, 2007, p. 2).

Context validity in Weir's framework means "the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (2005a, p. 19). Weir argues for the use of *context*, as it refers to the social dimension of the language. He also endorses Anastasi's (1988, p. 132) view, who pointed out that rather than item content, it is the relevance of the test taker's response in the given

context that need to be accounted for. It is not only the domain, skills, and language ability that count but also their operationalisation by the test taker. To be able to observe this, Cambridge ESOL uses an 'observation checklist' for their oral examinations (O'Sullivan et al., 2002).

In the process of designing writing tasks, context validity is of particular importance. To be able to test the underlying abilities which are necessary for writing, the task and its context has to be familiar and acceptable for the test taker. In other words, it has to be as authentic as possible in the domain of the test. For this reason, situational authenticity is part of context validity (Weir, 2005a, p. 56). In addition, the task requirements should be adequately spelled out in the task so that it facilitates the test taker's goal setting and monitoring. Weir also argues that task design has serious effects on writing performance as the input affects knowledge telling and knowledge transformation (p. 59). Shaw and Weir (2007, p. 64) in their summary of the aspects of context validity for writing identify three main categories: (a) task, (b) administration, (c) linguistic demands. Here I discuss the issues related to task development for an academic test, for this reason issues concerning administration are omitted.

In relation to the practicalities of task design, Weir (2005a) elaborates the importance of setting task demands that will fulfil the requirements of representativity, as used by Hamp-Lyons, (1991). He highlights the categories of genre, rhetorical task and patterns of exposition (p. 68). As the response format has a strong influence on test takers' performance (Shaw & Weir, 2007, pp. 66-90), tests are advised to include more than one task (at least two according to the Hungarian accreditation requirements), which require a range of response formats. The prompt of the task should be clearly worded so that it facilitates planning and monitoring. In case of unclear purpose, the task is likely to be misinterpreted by the test taker, which jeopardises task performance and scoring. This way, context and scoring validity are clearly linked.

As for the linguistic demands, the tasks have to be appropriate in terms of genre and discourse model in the construct of the test. Weir (2005a) considers the question of content knowledge by adopting Douglas (2000) who divides it to background knowledge and subject matter knowledge. Douglas argues that writing tasks that assess languages for specific purposes require an interaction between the language ability and the specific purposes content knowledge of the test taker, and also the test taker and the test task itself (p. 19). Following this argument, Weir (2005a) puts forward that the task design in terms

of content should always consider the students' age group, experience, whether the topic is biased against a group of students, or unsuitable because of causing distress. In order to facilitate knowledge transforming, the task must observe the social and discursive practices as well (Hyland, 2002, p. 69). In relation to writing task design, Weir points out that the "consequences of not specifying an addressee in a writing task or in a reading test including a text meant for a different discourse community than the test takers are obvious" (Weir, 2005a, p. 80).

As regards scoring validity in general, establishing the validity of the rating procedure is especially important. As scores are used to make important decisions, objective assessment is crucial for all stakeholders of a test. In case of EAP tests, universities regard test scores as guarantees about the test taker's future performance (Kane, 2016, p.75). When a certificate is used for admissions purposes, the generalizability of the test score means that the students with the given certificate is likely to perform well during their studies (Deygers et al., 2017; Hyland & Hamp-Lyons, 2002; Knoch et al., 2015; Ringwald, 2018). Validation research in connection with university entrance exams seems to be of considerable interest in international language testing research today. Hyland and Hamp-Lyons speak about the necessity for students to master the "right English, to succeed in learning their subjects" (2002, p. 2). Analysing exam performance in connection with university entrance is also a highly investigated topic among language testers. Deygers et al. (2017) look at how entrance exam scores reflect the extent students cope with the linguistic demands of their studies. Ringwald (2018) in her presentation at the 15th EALTA conference EAP Special Interest Group meeting, put forward the preparatory power of the German school leaving exam (*Abitur*). Her conclusions include the critique of the Abitur, as her findings suggest that students who passed the exam with distinction still possess low academic skills. The main issues in the literature are the question of reliable rating scales (Deygers & Van Gorp 2015; Harsch 2018), the predictive validity of academic language tests (Harsch et al. 2017; Knoch et al. 2015), and the design of alternative rating tools, such as a checklist instead of a rating scale (Kim 2011; Struthers et al. 2013; Lukácsi 2017; 2018; 2020).

Validity and testing for admissions purposes are closely connected, since social mobility is one of the explicit objectives of the CEFR (Council of Europe, 2001, p. 1). Validity theories describe the character of language testing as a social practice (McNamara, 2012, p. 565), thus the above objective of the CEFR needs to be endorsed by

exam providers. The evidencing of objective and unbiased marking is one major requirement expected by international tertiary education institutions as the language requirements for university entrance have a gatekeeping function (Extra et al., 2009; Nagy, 2000). There are many factors that influence the success of learning in a foreign language environment, such as cognitive and academic skills, literacy and language level, financial background, or other social factors. As foreign language level is only one of the numerous factors one needs to be successful in foreign language programs, it is important to design assessment tools that do not only provide reliable scores, but also have positive effects on the social and academic skills of the test takers. Language tests might have a positive effect on teaching practices and skills development and have an "impact on the career or life chances of individual test takers" (Taylor, 2005, p. 154). Since knowledge of English is regarded as a commodity that may be regarded as a means to prosperity (Cameron, 2000), it is of crucial importance that test takers and university applicants be aware of and able to perform the practices of academic discourse (Weninger & Khan, 2013).

Test developers and exam boards believe that the construct of a given test describes the particular language use as intended and a precise definition of the construct will result in fair testing for all the test takers. Shohamy (2001, p. 4) problematizes this view, and she puts forward that the need for exploration does not concern measurement and testing *per se*, but rather the use of the tests and its future implications. Davies (2003) claims that language tests also have a political nature. Tests are selected to meet certain needs in society and as a consequence, "testers cannot expect that their work will not have a political dimension" (p. 361). Therefore, test developers and test providers need to pay attention to the (political) context of use. The consequences of the use of tests has been part of language testing literature since the introduction of Messick's (1998) unified view of validity, and it appears as a separate concept among the 'new validities'. "Consequential validity addresses concerns that tests should not be used in ways that are biased, are unfair, or encourage the unjust treatment of certain individuals or groups of people" (Hamp-Lyons, 2003, p.166). However, voices of the different stakeholders of a test would desire more attention, especially those of the test takers', as tests have direct influence on their lives.

Although large-scale tests of international organisations (e.g. IELTS, Cambridge Assessment, Pearson) claim to test *international English* (Taylor, 2002), bias in tests in language tests towards non-native speakers coming from different cultures is still present.

Uysal (2010) in her critical review of the IELTS writing test she found consequential validity issues, such as not considering the test use in terms of test takers who are coming from various rhetorical and argumentative traditions. Freimuth (2017, pp. 166-167) gives an overview of studies into bias in the IELTS examination, and concludes that a considerable amount of cultural knowledge is expected, identified writing prompts as cultural concern, and through content analysis, found references to cultural objects and political/historical settings of the English speaking world. Hall (2010) also raises the question of international tests being truly international in the context of 'world rhetoric' (p. 325).

The purpose of tests and how they are used in different policies is considered to be part of test validity. Kane (2019) points out the strong relationship between validity and test interpretation and use. He suggested refining the argument-based approach of validity and stressed that a test designed for a certain use might not prove to be a valid measure in a different context. The interpretation-use model is believed to be particularly useful as sources of bias can be predicted or identified.

The need for language tests for work and study purposes, citizenship, and migration put forward the need for examining the real-world context of tests as part of their validity claim. O'Sullivan and Dunlea (2015) proposed a definition of localisation for the British Council, Aptis research team. Based on this, Aptis researchers use localisation "to refer to the ways in which particular test instruments are evaluated and […] adapted for use in particular contexts with particular populations to allow for particular decisions to be made (O'Sullivan & Dunlea, 2015, p. 7). The Manual of Aptis is based on the socio-cognitive approach and in it, five levels of localisation is proposed to promote development projects in different contexts and facilitate communication between the developers and the different stakeholders of the test. They stress the importance of test use and purpose and deny the existence of the 'one test fits all' idea.

Tests specifically required for university entrance act as gatekeepers, however future success is not easy to define. York et al. (2015) found that academic success is a heterogeneous, complex construct influenced by multiple variables and proposed a new model for its description. At the same time, it seems that teachers are not aware of the different purposes and uses of language, which enhances the gatekeeping function of foreign language knowledge (Moore, 2007) in a negative sense. If there is no clear relationship articulated between test scores and real-world use, a "validity chaos" occurs.

(Fulcher & Davidson, 2009, p. 125) Language exams clearly have a justifiable gatekeeping function in the context of university admissions, but it is important to stress that all test takers are entitled to equal opportunities, thus their cultural background and context have to be taken into account. The above ideas certainly can be used as an argument for a locally developed test of Academic English in Hungary.

Chapter 5: Qualitative and Quantitative Research Methods Used in the Test Development Process of the Euroexam Academic Test

Quantitative and qualitative research are not sharply distinguished from each other but rather could be represented on a single continuum, where we may observe several possible variations of the two methods (Mackey & Gass, 2005, pp. 2-5). Although the different methods of the two types of research complement each other well, and integrating qualitative and quantitative data into a single study is methodologically beneficial and recommended, only in the early 21st century did mixed-type research become widespread in applied linguistics and linguistic measurement (Tsushima, 2015, p. 107). Turner (2013), reviewing three major journals in the field (*Assessing Writing*, *Language Testing*, and *Language Assessment Quarterly*), found that studies published up to 2003 rarely referred to the term "mixed methods research" in their methodology. In the first part of Chapter 5, I provide an overview of the quantitative and qualitative methods used in test development and test validation studies, then I present my research questions and outline the design of the Euroexam EAP writing task validation research.

The literature on linguistic measurement is dominated by quantitative research methods. On the one hand, these methods are indispensable for validation and reliability in research-based test development processes, and on the other hand, they are well suited for statistical analysis of data collected in the linguistic measurement of large populations (Chapelle et al., 2008). One of the main aims of purely quantitative research is to test hypotheses, in which the process of data collection and analysis is understood as objective (Mackey & Gass, 2005). Quantitative data collection and analysis is particularly popular in the field of language measurement and assessment since large sample data are suitable for generalization, research can be replicated, and results can be verified (Mackey & Gass, 2005, pp. 43-46). Tsushima (2015) also notes that in the development process of large-scale, standardised tests, an orientation towards quantitative methods is discernible. This preference is based on the need for the following: (a) score-valid, score-reliable tests, (b) the generalizability of the results, and (c) appropriate inferences based on the observations (Tsushima, 2015, pp. 106-107).

Since the integration of Messick's (1989) unified concept of validity into language testing, it has been accepted that language testing is not independent from the social context and culture where it takes place. That understanding does have consequences on research methods as well. In his seminal paper, McNamara (2001) applies the notion of

Judith Butler's performativity (Butler, 1990) to language testing. Butler argues that gender is not created based on something 'inner', but instead it is constructed through 'performativity', i.e. it emerges as an accomplishment of socially regulated practices. Transferring Butler's ideas to the field of language testing, McNamara contends that we have to realise that language testing is also a practice and is similar to the performative approach to the construction of gender. McNamara further argues that it is often the case that testing itself "constructs the notion of language proficiency" McNamara, 2001, p. 339). The other important study is by Lazaraton and Taylor (2007) in which they highlight the limitations of statistical data and argue for the usefulness of quality data. As a result, in addition to research based on statistical data analysis, the use of qualitative research methods has been increasing over the last two decades. Language testing has become more learner-centred and pays more attention to contextualisation and test-taker characteristics.

5.1 Quantitative methods and reliability

The two major types quantitative research is traditionally divided into are (a) associational and (b) experimental (Mackey & Gass, 2005). Both types look at the relationship between different variables: the former aims at determining the strength of an existing relationship through correlation, and is not concerned with causation; at the same time, the latter aims at revealing a causal relationship by manipulating a dependent variable and comparing it to an independent one. Quantitative data may be gathered by not intervening into the process, only by looking at participants or respondents carrying out a task as they would normally do. However, the above-mentioned manipulation usually takes the form of some kind of a treatment one or more of the observed participants receive. The collected data is suitable for statistical analysis based on which research questions can be answered and hypotheses can be tested.

Regarding language tests, both test-takers and testers would like to make sure that the results of the test represent the actual ability of the test-taker that was intended to be measured by the test. In addition to this, testers would also like to see that the scores are not changing under different circumstances. Although high-stakes tests aim to control the different factors in the different testing administrations, it is impossible to claim that nothing affects the scores. For this reason, statistical procedures are to be used to estimate the reliability of the test (Bachman, 2004, pp. 153-155). Reliability thus refers to consistency, and it is important to highlight that it differs from validity. While the validity

of a test measures the extent to which the test is a valid measure of the ability we want to test, reliability in classical test theory is the combination of the test taker's true score and the error score (Hatch & Lazaraton, 1991, pp. 45-46). In other words, true score refers to what we want to measure, whereas error score refers to what we do not want to measure. The problem is that it is impossible to observe the test-taker's true score, what we normally have is their observed score. Based on this, Bachman (2004) refers to reliability as the correlation between two sets of parallel scores (p. 159).

Following from the idea of parallel tests, the reliability of a test can be measured through the test-retest method. When using this method to determine reliability, test takers are given the same test at two different points in time. By administering the same test twice, it is possible to compare the test scores of the same population and measure reliability based on the correlation between the two results. The method has its shortcomings due to the interval between the two administrations. The test takers of the two administrations are affected by the so-called practice effect and memory effect (Bachman, 1990, 2004; Mackey & Gass, 2005). The scores of the two test administrations cannot be meaningfully compared because on the one hand the test takers are usually in the process of learning, which means their knowledge is not the same throughout the process. On the other hand, they might also as remember the test questions and the correct answers, which clearly influences their results when taking the test for the second time.

In order to eliminate the shortcomings of testing and retesting, the equivalence of forms method can be used (Bachman, 1990; 2004). This method uses two versions of a test, which are administered to the same individuals, and a correlation coefficient is calculated (Mackey & Gass, 2005, p. 130). Similarly to the test-retest method, using the same population has its advantages. With two versions of the same test, it is possible to remove the memory effect, and the practice effect – provided the two forms of the test are taken right after one another. When it is not possible for the same group of test-takers to take two forms of the same test at one time, equally powerful statistical methods like the split-half procedure and Cronbach's alpha may be used. Split-half procedure is used on data divided into two halves, in other words, the test takers receive two scores for the two halves of the test. The method of halving the test scores is crucial because it is not always possible to divide a test into to equivalent halves, and the test items are rarely fully independent. To avoid these potential problems, a correction formula, called the Spearman-Brown prophecy formula can be used (Hatch & Lazaraton, 1991, pp. 535-536).

In case the test consist of a large number of items, it is possible to calculate reliability from a single test administration using test level statistics, such as the number of items, the mean, and the standard deviation to determine internal consistency (Mackey & Gass, 2005, p. 130). One of the best-known coefficients is Cronbach's α or internal consistency coefficient (Bachman, 2004, p.165). The *Accreditation Manual* of the Hungarian Accreditation Board for Foreign Language Exams (Educational Authority, 2019a) gives 0.75 as a reference point for a reliable test.

The key to ensure the reliability of human rating is a well-defined construct, which makes sure that raters are rating the same underlying abilities. Raters' reliability can be considered among different raters (inter-rater reliability) or within a single rater's performance (intra-rater reliability). Rater consistency should also be considered to determine whether a rater is harsh or lenient (Bukta, 2013, pp. 91-92). The easiest way to calculate rater agreement is by using a simple percentage, which shows the ratio of agreements in raters' decisions, or Spearman Rank Correlation Coefficient (Bachman, 1990). Although there are many ways of calculating interrater reliability, one of the easiest ways is through a simple percentage. This is the ratio of all coding agreements over the total number of coding decisions made by the raters. As for agreement given in percentages, 75% and above is generally considered acceptable, correlation, on the other hand, may be considered good above 60% (Mackey & Gass, 2005, p. 244). The shortcoming of these simple calculations is that they do not take the possibility of co-occurrence by chance. To make up for this, the formula to be used is Cohen's kappa (Cohen, 1960) or Krippendorff's alpha (Krippendorf, 2004). The acceptable range for both coefficients is above 0.8.

5.2 Qualitative methods

The research based on qualitative data collection is not numerical but is based on interpretation. These studies are typically open-ended, and are therefore suitable for exploring soft data, such as the cognitive processes, feelings, and impressions of informants, in my case raters, using the informal categories used in the process. The qualitative approach is critical in nature, i.e. the research is often created for a specific social or (educational) political purpose (Scollon, 2001, p. 139).

The term qualitative research refers to a set of different methods of data collection. For this reason, it does not have a uniform model, but there are different traditions where it

is rooted. Generally speaking, qualitative research is defined as research that uses descriptive data but does not use statistical procedures. Based on this definition, qualitative research can be described by several characteristics (Mackey & Gass, 2005, p. 162-165). These characteristics are as follows: (a) qualitative data has a descriptive nature; (b) researchers usually observe participants in their natural environments, or take their social context into consideration; (c) data collection does not aim at observing large groups, they prefer deeper, more intensive work with fewer participants; (d) the position of the researcher usually takes an emic, rather than an outsider, etic perspective; and finally, (e) qualitative researchers usually follow a cyclical, open ended process and let categories emerge from the context.

Data collected through the active cooperation of participants and introspection can take many different forms and require assistance by the researcher (Dörnyei, 2007, p. 124; Mackey & Gass, 2005, p. 77). In the field of language testing, both test-takers' and raters' internal processes, such as thoughts, feelings and motives are essential to integrate into research. In order to explore these characteristics all, the following ways of soft data collection and analysis may be used: case studies, interviews, verbal protocols.

Case studies are often used in applied linguistics to explore people, communities, and contexts. A case may also refer to revealing a single person's motivations and dispositions. For this reason, case studies are rarely generalizable, but through getting to know the context, the collected data is usually very detailed and highlight the steps of processes (Mackey & Gass, 2005, p. 172). Although the generalizability of case studies is an issue in research design, it is possible to draw conclusions based on data provided by a small sample, provided the field of study is not yet explored. For this reason, following purposeful sampling, and combined with other research methods, the results may be accepted as valid and generalizable (Dörnyei, 2007, p. 153-155).

Another frequently used data collection method of qualitative research methodology is conducting interviews. Research interviews are divided into three different types based on their degree of structuredness. The most standardised interview type is called structured interview, in which the researcher uses the same questions with all the respondents. This type of interview is very similar to a written questionnaire. The so-called semi-structured interviews are less standardised, the researchers form only a few key questions in advance and in the process of their attentive listening they may ask new ones and often adapt the original ones to the situation and the respondent's answers. The third

type, the so-called unstructured interviews are the most ‘natural’ of the three in that they are very close to everyday informal conversations, which are shaped by both parties rather than driven by the interviewer’s ideas (Mackey & Gass, 2005, pp. 173-175). In addition to one-to-one interviews, focus-group interviews are also widely used in research, where a group of people discuss a main topic (Dörnyei, 2007, p. 137). The interactive verbal nature of interviews allows for researchers to gather data that would be hidden in a rigid, predetermined (written) format. Respondents often feel more comfortable and open in an interview situation, and their answers are usually characterised by verbosity.

In the course of item development, cognitive processes of test-takers and raters are essential to be revealed, in order to see what knowledge, skills they use for answering the test items (Paulsen & Levine, 1999, p. 4). It is important to realise that this information usually remains hidden when quantitative methods are used but it is a crucial part of the development process as it enhances the reliability and validity of language tests. This aspect of the process can be explored with the help of verbal protocols. Verbal protocols are also referred to as introspective interviews (Dörnyei, 2007, p. 147). As for the timing of the interview, the protocols can be retrospective – also referred to as immediate or stimulated recall; or concurrent – in other words, think aloud protocol (Gass & Mackey, 2000, p. 13). Verbal protocol analysis is often applied in research connected to test development. In retrospective protocols, respondents recall what they were doing after completing a task, whereas in concurrent protocols, they are commenting on a task while performing it. The main stages of the analysis are: (a) data collection, (b) coding, and (c) data analysis (Green, 1998, pp. 1-4). The coding stage, in other words, transcribing the verbal protocols is of importance in the process because replicability is key for qualitative research methods as well (Gass & Mackey, 2000, p. 3).

Verbal protocol analysis has its advantages and disadvantages. Dörnyei (2007, pp. 150-151) suggests that in general, the method is advantageous because it gives insight to hidden thoughts and feelings; if properly designed and conducted, give access to cognitive processing in the course of task completion. On the other hand, we also have to note that there we might face information loss especially in the case of retrospective interviews; moreover, thinking aloud while performing a task might hinder task completion.

5.3 The advantages of mixed-methods research

Both quantitative and qualitative methods have their advantages and disadvantages. Research-based test development studies generally prefer using both methods to enhance their validity claims. Combining the two methods is referred to as mixed-methods research (MMR). Jang et al. (2014, p. 129) also highlight that MMR is not only a mixture of methods, but it is an extension of the methods of enquiry. Dörnyei (2007, p. 164) summarizes the advantage of MMR in two points: (a) gives a fuller understanding of a given phenomenon, and (b) gives the possibility of verifying one set of data against another.

Following from the above, the greatest strength of MMR is triangulation, which allows the cross-validation of findings through the use of different methods. Mackey and Gass (2005, p. 181.) identify different types of triangulation: (a) theoretical triangulation (using multiple perspectives to analyse the same set of data), (b) investigator triangulation (using multiple observers or interviewers), and (c) methodological triangulation (using different measures or research methods to investigate a particular phenomenon). In case one method is not sufficient to provide ample support for a claim, a number of independent sources are to be used to support the findings.

In a validation study, both qualitative and quantitative evidence should be collected. Quantitative approaches include (a) measuring reliability, and (b) describing relationships, using Classical Test Theory (CTT). Qualitative data collection techniques, at the same time, include (a) case studies (Mackey & Gass, 2005), (b) verbal protocols and (c) test taker observations (Council of Europe, 2009). Collecting and analysing quantitative and qualitative data in a single study and integrating it into several phases of scientific research is a priority (Creswell et al., 2003, p. 212), with particular emphasis on triangulation, which is used in cross-modal research to cross-check findings, and to counteract the distortions associated with the various methods. Thus, in mixed-methods research, the different methods can be considered complementary to each other, since the analysis and explanation of the results of one method can be supplemented and compared with the results of one another.

The relevance of using mixed-methods is ultimately supported by the basic argument about test validation within the argument-based approach. It contends that validity is never the property of the test itself but “it is a function of the way in which the

results can be meaningfully interpreted as measures of the underlying construct, when the test is administered to a specified population of test takers" (Paltridge & Phakiti, 2015, p. 473). It also contends that gathering evidence-based data through a well-designed research procedure is in the interest of all the stakeholders of a high-stakes language test (Yin, 2011). Therefore, it does not suffice to claim that certain tasks measure academic abilities just because they were designed to look like tasks from other academic tests. If we want to claim that a test is valid, i.e. it measures what it is supposed to measure, we need to argue for the validity of the test, and produce a sound theoretical reasoning that builds upon various kinds of empirical evidence (Chapelle et al. 2008; Read, 2015; Shaw & Weir, 2007; Weir, 2005a). Argument-based approaches to test validation and evidence are needed to be adapted for specific test purposes, constructs and task/item formats, and also for new modes of delivery (Fulcher, 2010; Fulcher & Davidson, 2007). I presented the concepts of validity and validation research in the first part of my dissertation; in the following sections of this chapter, I demonstrate how they relate to my specific topic and research design.

5.4 Test development and validation

The aim of the dissertation is to build a validity argument about how the construct of the proposed writing tasks in an Academic test reflects the skills required in higher education, and whether the results reflect reliable scores and unbiased marking. The method is built upon Weir's (2005a) theoretical framework and the characteristics of test usefulness (Bachman & Palmer, 1996; 2010), and consider Read's (2015) validation stages, using a mixed-methods approach.

As discussed in the literature review, Bachman & Palmer (1996; 2010) introduced a model of test usefulness with six characteristics to consider in test development. In order to end up with a valid test, one must review the (a) reliability, (b) the construct validity, (c) the authenticity, (d) the interactivity, (e) the impact and (f) the practicality of the test tasks. Although Bachman and Palmer's work is still influential in test design, later Bachman (2005) pointed out that these categories are alone standing, and the relationship between them is not clearly defined. The only thing we can do is that we build "a convincing case that the decisions we make are defensible and supporting that case with credible evidence are the two components of the validation process" (Bachman, 2005, p. 5). When considering the six characteristics, we have to realise that it is impossible to achieve a high

quality for all the characteristics, there are certain compromises we have to make, and instead of the 'perfect test', we have to focus on designing a 'good enough' test. Out of the six characteristics, the ones which are the most important for the purposes of EAP test design – to see how well the test tasks fit into the academic context and discourse (Chan, 2013) – are (a) reliability, (b) construct validity and (c) authenticity. It is important to design test tasks which provide comparable results in different administrations, measure what we want to measure and are representatives of target language use.

These three characteristics of Bachman and Palmer (1996) appear in Weir's (2005a) framework as the components of validity. As regards construct validity, Bachman and Palmer claim that it is essential that the construct is valid in a specific context. This idea was further developed by Weir (2005a), who uses construct validity as an umbrella term and introduces new aspects of validity. In his framework, the construct is determined by the context, and authenticity appears as an integral part of context validity. In case of a writing task, context validity is about mapping the linguistic and content demands of a test task, and the demands of the real-life writing tasks in the target language, i.e. we have to see whether we are testing target language use in a specific context. He also introduces scoring validity, i.e. the validity of the rating procedure in which he integrated the notion of reliability (Shaw & Weir 2007; Weir, 2005a).

The main focus of the present research-based test development process is generating validity evidence for the writing tasks of the locally developed Euroexam Academic test. I used the theoretical framework of Weir (2005a) to build up the different stages of validation that the tasks had to go through. First and foremost, to establish the context validity of the test tasks, we have to reveal whether the proposed test tasks (Task 1: *formal transactional email* and Task 2: *discussion essay*) are representative of the target language use in an academic context. Following that, to establish the scoring validity of the tasks, we need evidence about how test takers and raters approach the task.

The different types of validities outlined above are relevant – to a different extent – for the two text types of the writing paper. As for Task 1, there is a need for evidence that transactional writing is part of the academic domain, whereas the question I identified in connection with Task 2 is in connection with the quality of the essay. Thus the aim of the investigation in connection with transactional emails is to reveal whether they are part of the academic domain, but as for its assessment, the dissertation does not look into how the content and structure has an impact on the assessment and how the established scoring

validity may be improved. The validation stages (Table 2) I designed for the research-based development and validation process of the Euroexam Academic focus on the two tasks differently. There are two underlying reasons for this. Firstly, this is a feasibility question. The dissertation itself has a more limited scope than the entire validation process. It is possible to provide evidence for the scoring validity of Task 1 and the context validity of Task 2 based on the standard stages of validation, such as external expert judgement, verbal protocols, trialling and pretesting (Weir, 2005a). In addition to this, it is important to point out that the validation process is within the context of the Euroexam writing tasks, which means that the two tasks of the newly developed test have to be in line with the Euroexam writing construct and the Euroexam portfolio.

5.5 Research hypotheses and research questions

The aim of the dissertation is to establish a validity argument for the writing tasks of the Euroexam English for Academic Purposes Test, with special attention to the context validity of Task 1 and an improved scoring validity of Task 2. The research questions are as follows:

Research Question 1: Is transactional writing a valid task type for an EAP test?

Apart from discursive and argumentative writing, which appear both as authentic tasks in university education and in EAP tests, the main question concerns the validity of transactional writing in a test for Academic English. The research hypothesis implies that transactional writing is also part of students' repertoire. In addition to the professional side of academic life, university students are expected to arrange their studies, and develop and nurture issues in relation to administration and registration. Apart from meeting academic requirements, students are expected to meet the demands of formal communication regarding their studies. Based on this assumption, transactional writing is also part of the academic domain, therefore formal transactional text types are what students most often write in an academic context.

Chapter 6 of the dissertation investigates this question through empirical research and expert judgement. As part of the domain analysis a small-scale preliminary study was carried out to investigate the construct with the following secondary research questions:

a) What are the most frequent written genres regarding communication between university undergraduates and members of staff?

b) Is formal written communication in English a part of university students' target language use (TLU)?
c) How important is the level of formality in TLU?

These questions aimed to disclose whether the proposed transactional writing task is suitable for an Academic exam using qualitative methods. The results of the preliminary investigation served as a basis for preliminary task design and the secondary research questions were addressed in the questionnaire used for expert judgement.

As regards scoring validity, the research questions are based on both quantitative and qualitative enquiry. The results of the verbal protocols and statistical analyses in Chapter 8 try to reveal the advantages of checklist-based assessment. The research hypothesis proposes that a task and level specific checklist-based assessment tool improves the objectivity and reliability of the assessment of Task 2. The hypothesis is tested through the following research questions:

Research Question 2: Compared with a marking scale, can checklist-based assessment enhance

- the objective scoring of academic discussion essays and
- rater reliability?

The secondary research questions addressed in the course of the analysis using Classical Test Theory are as follows:

a) Is the reliability (Cronbach's alpha) of checklist scores high enough to fulfil accreditation requirements?
b) How do checklist items perform in terms of item difficulty and item quality?
c) Is the checklist capable of discriminating low and high performers?
d) Does checklist-based rating affect the success rate of the essay task?

Test scores are of particular importance for the different stakeholders of a test – universities, awarding bodies, test takers and raters. The main issue with the rating procedure of Euroexam is that test takers and raters have different perceptions of what counts as successful writing performance (Lukácsi, 2017). In addition to this, ratings may be subject to personal judgements and halo effect (Knoch, 2009), even "trained experienced raters have been shown to differ systematically in their interpretation of routinely-used scoring criteria" (Eckes, 2009, p. 5). Previous research at Euroexam

International (Lukácsi, 2017; 2018, 2020) proved that a level and genre specific checklist enhances the objectivity and reliability of scoring a B2 level transactional writing task.

The verbal protocols with Euroexam raters in Chapter 7 were aimed to reveal how the raters approach the essay task during scoring. The verbal protocols also shed light on how the features they associate with a well-formed essay differ from each other. An additional qualitative enquiry in connection with scoring validity concerns Euroexam raters' ideas about the writing product:

Research Question 3: Can checklist-based marking increase the genre awareness of raters?

Based on the teacher verbal protocols and the rater think-aloud protocols in Chapter 7, a checklist-based rating tool is designed based on dichotomous statements and concept check questions. Throughout the development of the level and task specific checklist, teachers' and raters' verbal protocols serve as a basis for qualitative analysis to design a checklist that may guide raters towards a common understanding of the genre of the essay.

5.5.1 The context of the Euroexam Academic test

The Euroexam academic validation process cannot be viewed without its context. The locally developed test has to be integrated into the Euroexam portfolio and the system of state accredited language exams in Hungary. Euroexam, being a Budapest-based company, according to Government Decree 137/2008 (V. 16.), offers state accredited general and business language test at B1, B2 and C1 levels in Hungary. In addition to this, the company is present in the international market, where they provide language tests from A1 to C1 level worldwide. The Euroexam C1 level tests gained UK Naric recognition in 2017. Both the Hungarian accreditation and the UK Naric board confirmed that Euroexam offers language tests that are objective and valid measures of test takers' English language ability. As the tests of Euroexam are highly standardised tests, the development procedure of test tasks follows a manual (Euroexam, 2018c) to make sure that the sets of tests are parallel and test the same construct. The research aiming at establishing the validity of a new test consequently has to be placed in this context. Framing the research might seem to be a limitation, however it is not against the basic ideas of a valid test and validation research (Bachman & Palmer, 1996; O'Sullivan & Dunlea, 2015; Weir, 2005a).

The evidencing of objective and unbiased marking is one major requirement expected by international tertiary education institutions as the language requirements for

university entrance have a gatekeeping function (Nagy, 2000). To make sure that applicants possess the skills based on their language certificates, it is important to design an assessment tool that does not only provide reliable scores, but also has a positive effect on the skills of the test takers. Language tests might have a positive effect on teaching practices and skills development and have an "impact on the career or life chances of individual test takers" (Taylor, 2005, p. 154). Since the knowledge of English is regarded as a commodity (Cameron, 2000), it is important that the test takers and university applicants be aware of the practices of the academic discourse (Weninger & Khan, 2013).

The marking procedure of writing tasks has always been an issue generating interest in language testing research. The scoring validity of the subjectively marked writing tasks – especially that of the academic discussion essay – is a key issue in this regard to examine. The tool most assessment related handbooks describe for the assessment of writing products is the rating scale (Alderson et al., 1995; Bachman & Palmer, 1996; McNamara, 1996; Shaw & Weir, 2007; Weigle, 2002, Weir, 2005a). At the same time, the fallacy of subjective marking of learners' writing performance and the need for more objective, i.e. consistent assessment has been repeatedly raised by a number of publications (Eckes, 2009; Knoch, 2009; Knoch, 2011, Lukácsi, 2017; 2018; 2020; McNamara, 2000) as well as Chapter 9 of the *Common European Framework of Reference for Languages* (Council of Europe, 2001).

The necessity of an objective rating tool for the Academic test is twofold. On the one hand, as it was outlined in the literature review, the suitable rating tool for level testing is a checklist; on the other hand, scoring validity is especially important in connection with a high-stakes test that also serves as a proof of English language proficiency for higher education institutions. Further to this, the need for the creation of a new rating tool is rooted in the context. Since the number of international test takers started to grow in the past few years, there has been internal motivation for a fairer rating system from the company management.

The use of an objective rating tool is expected to reduce differences among raters and increase their genre awareness. Apart from this immediate result, there is a predicted positive washback effect that will develop students' genre awareness and writing skills and also increase the probability of the correct perception of their writing results.

5.5.2 The outline of the stages and methods of investigation

The dissertation handles the question of validity for the two tasks in two different ways. The qualitative and quantitative parts of the research may be regarded as complementary, the method of mixing shows a sequential structure, i.e. the research shows an iterative structure in which results and conclusions of each stage are built in the design of the following stages (Creswell, 2009, p. 14). The four stages of validation and the processes I have specifically designed for the writing tasks of Euroexam Academic test are displayed in Table 2. The steps of the stages of the development process were taken with regard to the writing construct of the Euroexam general C1 test. That is to say, both task types are examined in the first three stages, which are the standard stages of validation; however, the foci of the validation process are different for the two tasks. The approach of the present research-based validation process uses construct validity as a feature to unify the arguments (Kane, 2013), the present validity argument is based both on theoretical and empirical evidence, where the different validities (context and scoring validity) are linked through their interaction (Shaw & Weir, 2007).

Table 2
Euroexam Academic Test Writing Tasks - Stages in Validation

Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
Task 1 & Task 2	**Task 1 & Task 2**	**Task 1 & Task 2**	**Task 2**
Initial development	**Completion of test specifications and items**	**Pretesting test tasks**	**Establishing an improved scoring validity of checklist-based marking for essays**
Planning	Domain modelling and trialling	Evidence based analysis of test taker performance	Development of checklist items and CCQs
Domain analysis Preliminary investigation of the construct	Test taker characteristics	Student questionnaires	Verbal protocols
Expert judgement	Student and Rater interviews	Statistical analysis of pretest results	Rater and Candidate performance analysis

Stages 1 to 3 are the standard stages of validation research. Although I defined the focus and the main issues of my research in connection with the two tasks of the Academic test, I wanted the three standard stages to cover both tasks. The two tasks appear together in the initial development stage, trialling and pretesting stages. Stage 4 is an additional stage that was added to explore how an improved scoring validity of the essay task could be established. The first three stages of research-based validation follow the regular process

test providers go through when they design a new test. The novelty of the dissertation lies in the inclusion of a fourth stage. The aim of this stage is broader than the regular standard setting procedure. Suggestion for the improvement of scoring validity appears as an independent research project based on issues revealed in Stage 2 and Stage 3. The sequentially designed phases and the logic that links them are discussed in the data analysis.

In Stage 1, I carried out a small-scale preliminary investigation to find empirical evidence for the context validity of transactional writing in an academic test. The aim of the investigation was to define the construct. I carried out student and instructor interviews to see what requirements the students are expected to meet in the course of their studies, which served as a basis for the construct definition of the new Euroexam Academic test. As O'Sullivan (2012, p. 48-49) notes, this stage is rather informal, and the questions to be discussed vary between theoretical and practical. These questions help with producing the design statement, with which, by the end of this stage, a sketch of the test should be available.

Stage 2 is the core of the development phase, where I have to model the domain and turn to the test takers and raters for empirical data. The empirical research in this stage involves semi-structured student interviews to help me establish what is happening when candidates actually perform the test tasks together with illuminating "the cognitive processing that candidates go through in the test task" (Weir, 2005a, p. 233). In addition, in this stage I carried out rater think aloud protocols to see how Euroexam raters approach the task characteristics and assess test taker performance. It is important to highlight that domain modelling in case of test development cannot be separated from trialling the actual tasks. A small sample trial using strategies of qualitative inquiry helps determine the factors that affect task performance.

In Stage 3, after the small sample trial and the completion of test specifications, a larger-scale pretesting is administered; my aim was to reveal "how examination materials have worked in practice" (Weir, 2005a, p. 206). The sample size (N = 136) and the composition of the pretest, which are similar to the target population, allowed me to carry out statistical analyses and draw conclusions about how the items work under exam circumstances (Alderson et al., 1995; Fulcher & Davidson, 2007; Read 2015; Weir,

2005a). It is through this process that using the score data, standard setting usually takes place.

In Stage 4, I included an extra research enquiry and expanded the standard procedure of validation. This was necessary to achieve my ultimate objective, i.e. to improve the scoring validity of the current rating process of Euroexam. The main aim of the dissertation is to design a rating tool on the basis of which not only the results will be more suitable for university admissions purposes, but both test takers and raters will have a better understanding of the writing process and product.

The context and scoring validity of the two writing tasks may well be established through the first three stages, nevertheless, the existing experience regarding the assessment of the subjectively marked essay task calls for a further stage in which evidence for the validity of a new rating tool may be gathered. Evidently, there is need for validation research to create the new rating tool for both writing tasks in the Euroexam Academic test, and all the genres that appear in the writing paper of the Euroexam General English Test, however the dissertation only aims to analyse the genre of the discussion essay. I chose to focus on this particular genre for several reasons. First of all, I considered a practical perspective: the transactional email task is described in more detail in the specifications, moreover the guide for item writers specify the rubric and the input text for the task to a greater extent. This way, the construct determines the content and structure of the transactional email task, the specification describes in detail how long the text should be, it also gives the number and structure of content points and the two functions to be used. Furthermore, as regards the qualities of the two genres, we might say that compared to an essay, there is common understanding in connection with the content and structure of a transactional email. Consequently, the development project of the new rating tool in Stage 4 is focused only on the discussion essay. The scope of the dissertation not being wide enough to detail the development process of the checklist-based rating tool for the transactional email task does not mean that the exam centre is going to use two different rating tools for the tasks. The development of a checklist for Task 1 is set as a target for further research.

The above outlined four stages (Table 2) form three separate chapters of the dissertation (Chapter 6-8). As completing the specification and the test tasks, trialling the tasks and pretesting them logically belong together, I discuss Stage 2 and Stage 3 in one chapter (Chapter 7). The number of participants and methods of data collection of the four

stages are detailed in Table 3. The subjects of data collection will be high school and university teachers and students as well as experienced raters of Euroexam International, Budapest. As it is visible from Table 3, Stage 4 as an independent research project will take place between the pretesting of exam material and the live tests, and it will consist of eight phases.

Table 3
Outline of Research Methods and Procedure

Stages	Methods		Participants
Stage 1	Qualitative	Domain analysis	High school and university students, university teachers and members of staff (N = 11)
Stage 2		Student and teacher interviews	
		Rater interviews	Experienced accredited raters of Euroexam (N = 3)
Stage 3	Qualitative	Domain modelling verbal protocol:	
		a) test taker immediate recall	High school and university students (N = 6)
		b) rater think aloud	Experienced accredited raters of Euroexam (N = 3)
Stage 3	Quantitative	Pretest for testing material before it is used in live examinations – assessed with original rating scale)	High school students (N = 136) Convenience sampling
			Experienced accredited raters of Euroexam (N = 4)
		Candidate questionnaire for population analysis and statistical analysis	
Stage 4	Qualitative and Quantitative	1. Document analysis – CEFR, scoring tools (Hungarian, international context)	Experienced teachers of Academic English (N = 4)
			Stratified random sampling (non-proportionate)
		2. Teacher task completion: immediate recall	
		Create initial pool of items	
		3. Candidate performance marking: think-aloud protocols	3 scripts
		4. Reformulate and dichotomize binary choice decisions	
		5. Pilot 1	16 scripts
		6. Pilot 2	48 scripts
		7. Field testing	120 scripts

Stages	Methods	Participants
		Experienced and accredited raters of Euroexam (N = 9)
	8. Statistical analysis (Classical Test Theory)	

In Stage 4, I followed the methods used in the B2 transactional writing checklist development project (Lukácsi, 2018), and adapted them to the purposes of the C1 level essay checklist. Similarly to the previous stages of the research, the first phase of Stage 4 involved document analysis. It is important to review the CEFR (Council of Europe, 2001) and different scoring tools available both in Hungarian and international contexts. In order to identify the components of a good quality essay, experienced teachers who also act as raters took part in phases two to six, to find out more about both the components of the genre and its assessment. The four teachers took part in developing the items of the checklist through think aloud protocols during the rating process of writing products, interviews and carrying out assessment tasks with a different number of scripts. Altogether, I used 67 sample scripts in the development process. The last steps of the checklist development project entail producing and carrying out a field test involving 120 scripts and all raters of Euroexam International who are rating C1 level writing products. The aim of this phase is to see how the checklist works in assessing large live administration samples. This was done through statistical analysis of comparing means, standard deviations, result correlations with the two sets of scores, i.e. the original scales and the binary-choice items of the checklist.

The main aim of the dissertation is to establish the validity of the EAP writing tasks with special emphasis on establishing an improved validity of the rating procedure. The major contribution of the dissertation is the design of the checklist-based rating tool. The present research also involves the analysis of the genre awareness of raters, and an investigation of test takers' performance of the writing tasks and their results using a mixed methods approach. In this regard, the research is highly relevant for English instructors due to the positive impact and the washback effect the rating process might have on exam preparation courses.

Chapter 6: Initial Development

In this chapter, I report on the first, initial stage of the Euroexam Academic test development project. Stage 1 of the validation research focuses on the context validity of the two writing tasks proposed to be included in the Euroexam Academic test, namely Task 1: *formal transactional email* and Task 2: *discussion essay.*

As mentioned in Chapter 5, the validation process is built on Weir's (2005) theoretical framework and the stages were designed using Read's (2015) validation process. The importance of a well-designed initial development stage is twofold. On the one hand, exploring the theoretical background of foreign language skills and their use in different domains helps define the construct and build an adequate basis for the actual test. On the other hand, *a priori* validation (Weir, 2005a, p. 17) will contribute to the reliability of the instrument. Providers of high-stakes tests have to be able to show to the different stakeholders that they can trust the fairness of the testing process and the test results. In the name of that requirement, this chapter focuses on the considerations of the planning phase: a brief summary of the requirements of the Hungarian accreditation system and the Euroexam portfolio vis-à-vis those requirements; the description of the steps of domain analysis, the presentation of the preliminary investigation within the domain analysis concerning the context validity of the transactional email task, and finally the results of the expert judgement in relation to the two writing tasks.

6.1 Planning in the context of The Euroexam portfolio

Euroexam International, founded in 2000, operates mainly in the context of state accredited language exams in English and German in Hungary, but they also offer general and business language tests altogether in 80 countries. The general English tests are internationally recognised and thus they are operated at five levels of the CEFR (A1-C1), whereas the Hungarian accreditation refers to levels B1, B2 and C1. The English for Specific Purposes test (Euro Pro – Test of Business English) is available at B1, B2 and C1 levels. Due to the Hungarian accreditation requirements, the design and development procedure of assessment tasks needs to observe a number of well-defined criteria.

In addition to considering the existing regulations and suitable measures, awarding bodies and foreign language exam boards start a development project by outlining a research project for generating validity evidence for the test proposed. The validity argument in this phase has to be explored "in view of the fact that the test has not yet been

used" (Chapelle et al., 2008, p. 320). Therefore, researchers need to find external evidence to support and justify their validity argument. Language testers who are involved in large-scale assessment may rely on handbooks that are specifically designed to help test design and validation (ALTE, 2011; Lane et al., 2016; Newton, 2017; Weir, 2005a). In addition to the handbooks, there is a great number of research reports on validation procedures available (Chapelle et al., 2008; Newton & Baird, 2016; Read, 2015; Shaw & Crisp, 2012).

Despite the diverse studies investigating validity, consulting the relevant literature is not sufficient for generating validity claims in the course of a test development project. There might be similarities between various tests as regards their purpose or target population, but each test is different, therefore a bespoke validity argument needs to be produced for each test. At the same time, the local context and the existing portfolio can never be disregarded (O'Sullivan, 2018). For this reason, the Euroexam Academic development project did not start from scratch. Fitting into the portfolio was also a specific requirement from the management of Euroexam International, who imagined the Euroexam Academic test as a specific new profile extension. Extending an existing construct with a new profile may make a development project shorter and simpler, all the more so because the financial burden of the accreditation process is significantly smaller (Educational Authority, 2019).

Euroexam International set up an Academic Development Team, the members of which were assigned different roles in the development and validation process. Each of the four skills to be tested was allocated to a member of the team. My role and responsibility in the team was leading the research-based development process of the tasks of the Writing Paper. I observed the characteristics of the writing tasks of the existing tests in the Euroexam portfolio and listed the relevant ones for the new academic test. I also consulted the *Accreditation Manual* (Educational Authority, 2019a) to see what changes can be made that result in a new test profile but maintain the existing construct of the test. Based on this review, I decided that the new test should follow the existing practice in three aspects: (a) task design, (b) construct relevance, and (c) skills to be tested. As for task design, to make sure that the ideal L2 knowledge is reflected in the construct defined earlier, the tasks of the new test is designed to be as similar as possible to the other ones in the portfolio, and the test taker is tested with the same number of tasks. The skills to be tested also relate to construct relevance, the new test is testing the fours skills (Reading, Writing, Listening and Speaking). The accreditation requirements in Hungary also prescribe a separate testing of

these four skills, which means that integrated tasks, however authentic they are in academic tests, cannot be used.

Each time a new test is being developed, the claims need evidence and support, but as Kane (2016) puts it, "we do not need to reinvent the wheel for each validation effort" (p. 78). Validity, therefore, has to be interpreted within the context of the Euroexam portfolio, and the valid text and task types have to fit in the context of Hungarian state accredited language exams. The test tasks should be suitable for level testing, which means that test papers at each level are aligned to the CEFR. The Euroexam writing construct is specified in the *Euroexam Detailed Specifications* (Euroexam International, 2019b) document. The Specifications claim that "adopting a socio-cognitive approach, the general description of writing tasks comprises (a) domain, (b) content knowledge, (c) cognitive processing, (d) instructions, (e) L2 proficiency, and (f) authenticity" (p. 23). The C1 level specifications (Euroexam International, 2019a) and the *Euroexam Guide for Item Writers* (Euroexam, 2018c) both specify in detail what variables (context, linguistic, discourse, etc.) appear in the framework. Based on these two documents and Lukácsi's (2013) review of the B2 level specifications, we may conclude that the Euroexam writing construct uses the Grabe and Kaplan model (1996).

According to the *Euroexam Detailed Specifications* (Euroexam International, 2019b), the skill focus of the two writing tasks in question is different. Task 1, the formal email is transactional, whereas Task 2, the discussion essay is discursive writing. In Task 1, test takers respond to an input text and produce a formal response for an intended recipient; in Task 2, test takers write an extended text that is guided by a short prompt, inviting them to mobilize their relevant knowledge and experience. The word count for both tasks is 200-250 words, and test takers are given 60 minutes to complete the two tasks. The tasks are to be completed on paper by ink, the use of dictionaries is allowed. Task 1 is obligatory for all test takers while Task 2 is included in a list of three kind of exercises of which the taste taker must choose only one. The layout requirements of the first task include a design that makes the task "as authentic as possible" (Euroexam International, 2018c, p. 25), which means that test takers are given multiple sources (maps, leaflets, timetables, emails) and content points as prompts. The specifications make it clear that the tasks should have a specific purpose beyond merely describing or comparing. They should use the given information to express an opinion, justify a request, explain a situation. It is important that the writing task have a purpose and the writer a sense of

audience. The other point in connection with the construct of second language writing is the importance of linguistic resources. It is highlighted in the specifications that in order to achieve the task successfully, the test taker needs to "demonstrate a command of English expected at this level", "also tasks should allow the candidate to demonstrate lexical and grammatical range and accuracy in a given area" (Euroexam International, 2018c p. 29), which again is a clear reflection of the linguistic resources described in the Grabe and Kaplan model (Grabe & Kaplan, 1996, pp. 220-221). The second task is an optional one, which means that the test takers can choose from three titles and genres. Instructions for each task are always textual and limited to 60 words, respectively. The genres test takers can choose from are the following: (online) article, review, essay, report. It is important to stress that the test takers are instructed to bear in mind the intended audience, but at the same time they have to use their own knowledge and experience to complete the tasks, i.e. content points or input texts are not given.

In order for the writing construct to remain unchanged, the tasks of the Writing Paper of the newly developed EAP test must meet the requirements specified in the *Euroexam Detailed Specifications* (Euroexam International, 2019b), and may differ only in their profile from the original. However, as the Academic test differs from the General English not only in its subject matter but also in its purpose and future use, it is necessary to examine whether Task 1, the *transactional writing* task and Task 2, the *discursive writing task* and its genres are acceptable and can be interpreted as valid in the Academic domain.

6.2 Domain analysis

The aim of domain analysis is to gather information and evidence about the domain of the test, including concepts, representational forms, social knowledge and interaction (Mislevy & Riconscente, 2005, p. 7). Regarding the design of language tests, the findings help researchers in making claims about the target population and the target language use (Bachman & Palmer, 1996). Normally, the domain analysis phase is looking for underlying characteristics and knowledge in a specific domain, and researchers do not usually have concrete tasks in mind, they rather set out to generate validity claims for future tasks of the assessment. In the case of the Euroexam Academic Test, however, the process of domain analysis followed an inverse structure. When I set out to explore the domain of academic English as the target language use of students in tertiary education, I specifically wanted to

reveal whether the two tasks of the General English test I used as a guideline could be seen as valid task types in an academic language test. The reason for this was the need to conform to the construct of the existing writing test. As the first step of domain analysis, I reviewed the relevant literature on the basis of which I considered the specific requirements of state accreditation in Hungary; then I looked at the task types that are used for Academic tests and compiled a short overview of academic language tests that are available in Hungary.

The first step I performed in domain analysis was to find out what tasks are used in real life in the academic domain (Chan, 2013, p. 49) and how they can be adapted for testing purposes. Real-life tasks and their use in a testing environment are connected to the question of context validity and authenticity. McNamara (1996, p. 11) points out that test tasks are most of the time simplifications of real-life tasks, and the testing situation is always artificial. As an example for these concerns, Weigle (2002, p. 52) problematizes the task type of impromptu essay in academic writing tests. She claims that although essays are authentic text types in comparison with real-life academic writing, an impromptu essay is yet far from being authentic for three reasons. Firstly, no source materials are provided, so the test taker has to write from knowledge and experience, secondly, real-life academic writing is rarely timed, and thirdly, the assessment of real-life academic writing focuses on content rather than language and organisation. We may conclude that the notion of authenticity will always be relative to other potential task types. The independent writing task is much more authentic than a multiple-choice writing task, especially if we consider the definition of authenticity in terms of face validity, as formulated by Bachman (1991, p. 690), a test is authentic if it looks like a test. This view is supported by the empirical research of Lewkowicz (2000), who found that authenticity was not an important characteristic of a writing task for test takers. It seems that for test takers, a test is authentic if they perceive it as an assessment.

Another aspect one has to consider when examining real-life tasks for testing purposes is the adaptability of the real-life task to the testing situation. Although the requirement of practicality is usually connected to physical implementation, it has an effect on other aspects of writing assessment. It would be highly reliable and authentic if language tests could acquire multiple writing samples of test takers to assess their writing ability in different domains, but due to time constraints and test takers' limited attention span, the number of tasks that can be used to test writing ability in high-stakes tests varies

between two to four. The prevailing task types are independent and integrated tasks. In the following, I aim to review academic writing task types in the context of high-stakes language testing. EAP as it appears in a higher education context would be out of the scope of the dissertation. Therefore, real-life academic writing is only considered in relation to language test development for high-stakes language tests.

In the past decades, a number of studies reviewed the task types of academic language tests (Shaw & Weir, 2007; Weigle, 2002; Weir et al., 2013), and they all conclude that independent writing tasks, i.e. ones which do not require the use of reading sources, are prevailing in high-stakes language tests. The writing product is expected to be created based on knowledge and experience, and internal resources. The most common genres test takers need to be able to produce is a writing product based on a single line instruction, such as report, review, and argumentative essay (see task types of high-stakes EAP tests in detail in Table 4). Although the independent writing tasks outnumber the integrated task type in testing, in the language testing literature there has been a growing concern about the suitability of such tasks in testing academic writing (Cumming, 2013; Moore & Morton, 2005; Plakans, 2008; Weigle, 2002).

Based on the models of writing reviewed in Chapter 2, it can be concluded that academic writing requires writing expertise, and is considered to be a recursive knowledge transforming process (Bereiter & Scardamalia, 1987; Weigle, 2002; Weir et al., 2013). Knowledge transforming, apart from being a more complex cognitive process, might be understood as the use of sources. In EAP focused language testing, knowledge transforming skills of test takers may be tested through integrated reading into writing tasks. Gebril (2009) reviews the literature on academic writing models and gives a critique of the independent, timed, impromptu writing task. Chan (2013) is also in favour of the integrated task, she argues that "it seems inaccurate and inadequate to consider academic writing only as a productive language skill" (p. 22). The independent writing task has been criticised for its lack of authenticity, unsatisfactory cognitive validity, and issues has been raised in connection with the background knowledge effect and test fairness (Chan, 2013; Gebril, 2009).

Integrated 'reading-into-writing' tasks test candidates based on their response to reading sources (Weigle, 2004). Typical integrated tasks are discussion essays based on multiple reading sources, or report writing, and literature review based on input texts (Chan, 2013). As for context validity and authenticity, integrated tasks are considered

better than writing only or independent writing tasks (Cumming, 2014; Cumming et al., 2005; Plakans, 2012; Weigle, 2004; Weir et al., 2013). By using an integrated task in a high-stakes language test, test fairness can also be enhanced as it provides the test takers with equal amount of input and information on a topic. On the other hand, integrated tasks might be considered "muddied measurement" (Weir, 2005a, p. 101), since the definition of the construct could be an issue (Hirvela, 2004, pp. 43-45). The integrated tasks are muddy in that it is difficult to separate the constructs of reading and writing, and thus it becomes challenging for the raters to assess the writing skills of the test taker. It is the case because a difficult input text might hinder the test taker writing ability and weaker test takers might borrow more from source texts (Shi, 2004). A study of British Council by Moore (2015) investigated what would serve the purpose of a valid EAP writing task for Japanese university admissions. The results of the study showed that the suitable task for university entrance tests are direct, impromptu writing tasks, as they are suitable for the demonstration of a number of language functions and the knowledge of different discourse types. Furthermore, integrated tasks are out of question for another reason: they are not accepted according to the Hungarian accreditation requirements (Government Decree 137/2008 (V. 16.)) as they state that the four skills have to be assessed separately.

The academic writing genres of essay, report, summary, library research paper are the ones that have been found typically occurring in real-life academic contexts of learners (Carson, 2001; Cooper & Bikowski, 2007; Hale et al., 1996) and they also appear in the field of testing writing skills. Moreover, the CEFR (Council of Europe, 2001) suggests that a test should focus on what students can achieve with the language, and how they use different functions in different contexts for communicating concepts and ideas. Therefore, formal and informal communication is part of both General and English for Specific Purposes (ESP) tests in the form of transactional letters and email correspondence. Students, who pursue university studies, typically need EAP skills, which qualify them for the challenges representative of academic institutions (Ferris & Hedgecock, 2005). However, when I reviewed the websites of major test providers in Hungary to determine what EAP exams include (IELTS Academic, PTE Academic, TOEFL iBT) (Table 4), we can see that they do not test transactional skills: that is, written communication is not considered to be equally important part of the domain.

Table 4
Task Descriptions of EAP Tests Available in Hungary

Academic exam	Task name	Task description
IELTS Academic	Task 1	Test takers are presented with a graph, table, chart or diagram and asked to describe, summarize or explain the information in their own words. They may be asked to describe and explain data, describe the stages of a process, how something works or describe an object or event.
	Task 2	Test takers are asked to write an essay in response to a point of view, argument or problem.
PTE Academic	Summarise Written Text	After reading a text, test takers write a one-sentence summary of the passage.
	Essay	Test takers write a 200–300-word essay on a given topic
TOEFL iBT	Integrated Writing Task - Reading/Listening/Writing	Test takers write essay responses based on reading and listening tasks.
	Independent task	Test takers have to write from knowledge and experience and support an opinion in writing.

Although transactional writing is not part of the academic domain as defined by major test providers, it does not mean that it is legitimized by the lack of formal transactional writing in university students' life. Therefore, I wanted to establish if that is a salient genre in the academic context, and to what extent it is part of students' repertoire – if so, I could argue that it is a legitimate objective for me to propose the two tasks. Although the transactional text type (the formal e-mail in Task 1) is not typical in academic exams, my proposition is still that students have to be able to meet the demands of formal communication in connection with their studies on a daily basis. In general, people need to have knowledge about the different discourse types they have to take part in. Students need academic skills as well as social skills in university education (Hyland & Hamp-Lyons, 2002). One of the aims of my empirical research therefore is to gather data for the context validity of the proposed text type for Task 1, which is currently not part of EAP exams.

6.3 Preliminary investigation of the construct

As domain analysis always has to be evidence centred "to ascertain the kinds of tasks [are] appropriate for assessment" (Mislevy & Riconscente, 2005, p. 9), I decided to design a

small scale study to collect empirical evidence from the stakeholders in higher education in order to generate validity evidence for Task 1, a transactional writing task in an EAP test (Fűköh, 2018). I used the idea of context validity as presented in the framework of Shaw and Weir (2007), who interpret it as a mixture of linguistic and social and cultural demands; that is target language use in a specific social and cultural context. In addition to the professional side of academic life, university students are expected to organize their studies, and communicate with administration and registration. Students are expected to liaise with their tutors and university staff members. Although communication is almost exclusively electronic, it does not mean that there should not be a particular kind of formality of writing students have to meet. Based on this assumption, transactional writing is also part of the academic domain, therefore students often write formal transactional text types in an academic context. In order to test the hypothesis, the research questions I formulated for the empirical research as part of the domain analysis are the following:

a) What are the most frequent written genres regarding communication among university undergraduates?
b) Is formal written communication in English a part of university students' target language use (TLU)?
c) How important is the level of formality in TLU?

I approached university students as well as higher education faculty and staff to find out about their (electronic) written communication in relation to the emerging social and cultural needs. I aimed to reveal how these two well-defined groups in terms of their power position in the academic communication relate to the demands of transactional writing in an academic context. Two groups of participants took part in this case study. I approached a small group of university students (N = 5) and members of staff at European higher education institutions (N = 6). The students were all undergraduate international students, studying in Hungary, in two institutions, namely the Budapest Business School University of Applied Sciences and the University of Szeged. The students were recruited through instructors and the international coordinators of the two universities. All participants were BA students studying in Hungary, and all were recipients of an Erasmus or Stipendium Hungaricum scholarship. They were between 21 and 25 years of age, and their level of English based on self-assessment was B2. Their first languages were German, Russian (2 students), Serbian and Danish. The members of the staff group in the case study were academic and administrative staff from the Budapest Business School University of

Applied Sciences, Corvinus University, Budapest and King's College, London. The respondents were all in daily contact with international students, but worked in different positions (1 library assistant, 2 international coordinators and 3 university faculty).

The participants were provided relevant information regarding the background and purpose of the research via e-mail to meet up with me for an interview. Students were requested to provide sample emails they wrote over the course of the year spent studying abroad, while staff members were requested to bring with them for the interviews emails and messages they received from international students. Additionally, both students and university staff members were encouraged to draft current issues they experienced at the university. The research design included semi-structured interviews. The interviews were carried out individually in person and they were all audio-recorded and transcribed. The preliminary questions of the interviews can be found in Appendix 1. I conducted the interviews in English with the international students, and with the EAP instructor from the King's College, London, but used Hungarian with the other staff/faculty members. I provide their answers in my translation. In the transcription and coding of the interviews, to observe the research ethics and keep the identity of the speakers private, I refer to them as *Student* and *Staff member*, and distinguish the speakers only by assigning numbers to them.

The student and staff interviews altogether turned out to be a transcript of the 5000-word text that I analysed using MaxQDA software. In the course of the analysis, first I used colour coding to highlight the different themes, which helped me establish the thematic categories using the emerging topics in respondent texts. Through this coding process, the text was reduced to 1425 words. The content analysis of the text resulted in seven thematic categories provided in Table 5.

Table 5
Thematic Categories in Student and Staff Interviews

writing context	1. Purpose of writing 2. Topic of writing 3. Form of writing 4. Audience of writing
writing process	5. Practice of writers 6. Feedback from recipients 7. Formality

The seven categories can be grouped into two main areas. The first one is the writing context or rhetorical scene of writing comprising categories 1-4, purpose, topic, form and audience of writing (Carter, 2007; Connor, 2004; Hocks, 2003). The second area comprises categories 5-7, practice in writing a particular text, feedback on writing from the recipients, and the perceived level of formality that make up the writing process (Di Gennaro, 2006; Krapels, 1990), including practice in writing a particular text, feedback on writing from the recipients, and the two party's perceived level of formality.

The semi-structured interviews allowed for me to explore a number of topics, allowing the participants to freely elaborate on their writing experience (Given, 2008). The interviews offered an insight both into what and how students write in a university context outside their course requirements, i.e., what real-life tasks they identify for which they use written language, and also how the recipients' relate to their use of language and formality. After establishing the thematic categories, the emerging elements were identified in support of each category.

I start with the discussion of the four categories making up the context of writing in Table 5. All student participants spoke freely in response to my question concerning what and why they write relating to their studies. Based on their answers, the first category I could identify is the purpose of writing. In general, transactional features were salient in their answers: clarification, getting information, asking for explanation, explaining something and most importantly, arranging studies abroad. International students emphasized they use email correspondence extensively before starting their studies in a foreign country. They also said that drafting formal emails to accomplish the objectives is something new to this category of students, and often times, they do not typically possess neither the necessary skill sets, nor the experience regarding the drafting of formal correspondence.

As for the recipients of the transactional writing, the experience of staff and faculty members was of highly similar nature. International coordinators are usually the first to get in touch with international students. They confirmed that transactional emails are the most common form of correspondence. Based on their answers, the reason for the preference of written communication is twofold. On the one hand, students do not usually like oral communication through the phone because they easily feel intimidated; on the other hand, written answers can be referred to later on in the application process. At the same time, staff members, especially teachers highlighted that the written communication skills of

students very often lack professionalism, what they called “disorganised”, “clumsy”, or “lame”.

The purpose of writing (Category 1 in Table 5) with all students included arranging studies abroad. Studying in an unfamiliar environment involves an immensity of administration and organization (administrative issues, organizational issues). This supports the idea of using the function of getting things done within the academic environment. Seemingly, apart from professional genres, transactional features are clearly discernible in student writing.

The prompts of the interviews included questions regarding the form or channel of written communication (Category 3 in Table 5). Email was the most salient mentioned in the student group, all five of them identified it. (I present respondent texts using the MaxQDA coding system and layout.)

> **Text: Student_4**
> **Code: feedback on writing**
>
> …mainly email, I rarely use the phone or skype, I started emailing in connection with the studies

Additionally, all the answers included reasons for choosing this particular channel of communication: “easier, fast, available, efficient, and more straightforward”. International coordinators in staff listed email, skype and messenger among the communication channels through which they are available for students, while teachers in faculty mentioned emails only. Although staff and faculty members offer various channels of communication, the respondents highlighted that students usually prefer emails to other means, especially when they require confirmation or an official response regarding an issue.

Students also stressed that the addressee of the emails (Category 4 in Table 5) they write is nearly exclusively university staff (administrative staff, teachers, and library staff). The topic of private communication also emerged during the interviews, and students remarked that they generally prefer written communication with landlords and other students. Interestingly, email communication is less commonly used in this specific context. In their private life, they prefer instant messaging services, such as Messenger, Skype and Viber.

The three thematic categories (Category 5-7 in Table 5) that describe how writers approach a task and their perceived relationship to the context are broader than the topics

belonging to the context. These categories implicate the writers' understanding of their approach to a task and their perceived relationship to the context.

Regarding practice (Category 5 in Table 5), all participants said they never practised writing emails at school. Although a transactional email is a salient text type in everyday life and a typical language exam task, students reported no rehearsal or practice dating from their academic life.

> **Text: Student_1**
> **Code: feedback on writing**
>
> Absolutely no. Because in my time in Russian school I didn't get qualitative education.

> **Text: Student_3**
> **Code: feedback on writing**
>
> …at school no, I can't remember emails, I think we were taught to write letters. I used to write formal emails for my job, which helped me to develop my writing skills. No. I never practised anything like this. You have to know how to write an email.

This last comment has important implications regarding additional research characterizing the writing process, and it also links with teacher's feedback practices within the classroom. Earlier research in this field revealed that high school teachers in Hungary take students' writing skills for granted, therefore, they rarely offer feedback regarding the quality of writing (Molnár, 2009), which, strikingly, is similar to the remark of one of the respondents:

> **Text: Staff_member_2**
> **Code: feedback on writing**
>
> …students have to know how to write.

As regards feedback on writing outside the classroom (Category 6 in Table 5), I found that all students reported that teachers generally do not provide feedback on students' quality of writing outside the classroom:

> **Text: Student_1**
> **Code: feedback on writing**
>
> If I write some emails, I have never had critics. …in connection with my emails, no, never said anybody anything.

Only one student reported teacher feedback in connection with writing outside the classroom.

Text: Student_5
Code: feedback on writing

> Once I had a British teacher, and I emailed her to ask for a ppt stick for my presentation. She didn't say it in writing, but she talked to me personally in class. She told me that my request sounded rude and asked me to work on my communication skills. I didn't realize that I was being rude, she also said that the sentences were all correct, but the tone was not what she expected. It was a strange experience.

This experience conformed to the answers I got in the staff member interviews. Teachers and staff members are generally unwilling to offer feedback upon the style, register and language of writing. This practice is synonymous with the research of Knoch et al. (2015). It seems that there are two different attitudes towards feedback. Non-academic staff (librarian, office administrator) clearly put forward that they never provide feedback on the quality of writing, they restrict their communication to arranging things, resolving issues and clarifying matters. The answer of Staff-member 1 clearly states that providing feedback would cause negative feelings:

Text: Staff_member_1
Code: feedback on writing

> I'm used to all levels of language and various stylistic features. I do have concerns about how they write, but I don't think it would be polite to give negative feedback.

It seems, however, that the growing number of international students might generate more reaction to students' writing quality by faculty:

Text: Staff_member_4
Code: feedback on writing

> Outside class I don't usually provide feedback on writing quality, but now that I think of it… It happens that I take the time and in my reply I react to students' mistakes in their emails. I mean mistakes of communication, pragmatics…and students get back to me using the correct forms. It would be great to see a positive washback effect.

Staff member 4, who is a university EAP instructor, made it clear in the interview that they do not teach transactional genres in class and agreed that students are simply expected to possess the skills that enable effective communication without any instruction provided.

Regarding the question of formality (Category 7 in Table 5), staff members and students seem to associate the emails related to university issues to the genre of formal letters:

Text: Student_2
Code: formality

> Yes, letters and short essays. I used to write formal emails for my job. …my vocabulary is not enough for formal writing.

One of the student informants reported a case when the response she received was not as formal as she expected:

Text: Student_5
Code: formality

> I communicate very often with the international coordinator. He is very informal, I was surprised to see that he even included smileys, emojis in his emails. I also write to teachers, they are more formal, but it depends, you can see when they answer from their phones, these answers are much shorter and less formal. The external professors, lecturers are usually more informal than the professors of the university. But I think even if teachers are friendly, in writing you would be more formal with them.

Although electronic communication via e-mail must be fast, effective and straightforward, it is closer to formal letters. Chatting was only mentioned in connection with informal communication:

Text: Student_1
Code: formality

> We also have some chats in WhatsApp or Facebook, but it has less formal character. I mean that chat created by us and we communicate there in order to get more information about bureaucratic issues or some entertainment.

The student stressed that chat is something "created by us" in other words, students use chat to communicate amongst themselves. Students use chat both for entertainment and bureaucratic issues, thus, it is not the topic which defines the medium, but the audience. Students expect a certain level of formality throughout these emails, and they perceive them as part of formal communication, even if it exists only in the form of email.

6.4 Expert judgement

In the initial development stage, expert judgement was used to complete the information elicited through the interviews and to get further help in establishing the validity of the

new EAP test. According to Weir, “we need help from teachers and researchers in taking specific elements of the framework and determining their importance” (2005a, p. 214). For obtaining external expert judgement on the tasks I envisioned, I approached experienced language instructors and testing experts (N = 3) from Central European and UK institutions. I invited them to comment on the preliminary documentation of the test. I presented them with a preliminary test design of Task 1 and 2 and an initial description of the construct (Appendix 2) together with a questionnaire (Appendix 3). The preliminary test design, based on managerial decision following the accreditation requirements introduced above, followed the specification of the Euroexam General C1 test, the academic domain was only present in the topics of the tasks of the writing paper (See the list of academic topics in Appendix 4). The compulsory task (Task 1) was an example for a formal transactional email in the context of a university course and the optional tasks (Task 2) were examples for a review, an article, and an essay. The questionnaire I designed for the experts was of a multiple choice format and requested them to assess the two tasks in terms of the following six major points which were based on the characteristics of test usefulness (Bachman & Palmer, 1996, 2010): (a) the relationship of the test score and the construct, (b) the connection between the test task and the target language use, (c) the content of the target language use, (d) the features of the target language use, (e) the correspondence between contextual features and characteristics of the target language use, and finally (f) the test taker’s reliance on individual characteristics. For each question, they could indicate to what extent they think the material is adequate for the testing purpose. The options were: inadequate, limited and adequate. In addition, a text box offered them the opportunity to elaborate on their answers.

Based on their answers and the fact that they all opted for “adequate” for all aspects, I may conclude that the experts found the material valid overall, however, they all pointed out some shortcomings of the preliminary test material in terms of the target language use. Although the experts had not received the results of the preliminary study, they did not question the validity of the transactional writing task, but found the task type a valid example for target language use in the academic domain. They said that the test tasks are life-like and pointed out that the tasks: “measure what they wish to measure”, and “elicit from candidates the type of language referred to in illustrative descriptors mentioned in the relevant documents accompanied by the test”. They also called attention to the predictive validity of the tasks, i.e. the test takers “will need to perform very similar tasks

in their future studies." Expert 1 praised the idea of the transactional task in an academic test:

> **Text:** **Expert_1 – Task 1**
> **Question:** **b) Test task and target language use**
>
> International students need to do transactional writing, and this is neglected in other academic tests.

Although the context validity discursive writing was considered to have been established through literature review, and Task 2 was specified as using the typical genres of the academic domain (essay, report, review), the expert judgement raised concerns about using different genres as optional tasks in Task 2. As Expert 3 pointed out:

> **Text:** **Expert_3 – Task 2**
> **Question:** **b) Test task and target language use**
>
> Here my main reservation is whether the writing tasks are academic enough. It is in discursive writing that academic target language use differs more from general language needs, not transactional.

Expert 2 highlighted the importance of distinguishing between the general and the academic test:

> **Text:** **Expert_2 – Task 2**
> **Question:** **b) Test task and target language use**
>
> …the discursive writing required in academic settings (essays, reports and reviews) is quite distinctive. True, it one cannot fully replicate the task of writing a real academic essay within the time constraints of an exam, but as the test is specified one could complete Task 2 with a film review. In short, if you have to choose which task to make distinct between the general and the academic versions of the test, I would prioritise the discursive writing.

These ideas conform to the definition of academic language proficiency, which largely consists of the ability to cope with future studies. Proficiency in the academic field should be about testing the test taker's "ability to operate successfully in the English used in the academic domain" (Davies, 2007, pp. 84-85). Tests of Academic English should focus on the language of well-built arguments, analyses and explanations and should focus on all areas of the academic domain.

Although the invitation of external experts might seem as self-serving, their judgement plays an important role in domain analysis as they provide their evaluation

about the construct of the new exam. Based on their own experience and evoking the cognitive processes candidates go through during task performance, the expert questionnaires provided valuable practical and theoretical insight for me. Based on the results of the small-scale preliminary investigation and the analysis of expert judgement comments, I was able to redesign the two proposed tasks for the EAP test.

6.5 Conclusion

In relation to the **Research Question 1** (Is transactional writing a valid task type for an EAP test?) and its secondary research questions for the empirical study, Stage 1 of the development project, the preliminary investigation, and the results of the questionnaire for exploring expert judgement allowed me to draw the following conclusions:

a) What are the most frequent written genres regarding communication between university undergraduates and members of staff?

   The most frequent written genre among university undergraduates outside their classroom is the formal email. Apart from the academic genres, university students almost exclusively write formal transactional emails as a means of correspondence with academic staff in an academic context. Although they were not adequately prepared for this in school, students conclude, this is something they must know.

b) Is formal written communication in English a part of university students' target language use (TLU)?

   University undergraduates, academic staff interviews revealed that they regard formal transactional emails as part of their academic life and thus formal written communication is part of their language use.

c) How important is the level of formality in TLU?

   Students perceive formality as an important part of written communication generally preferred within this setting, and in relation to the particular audience.

Based on the answers to my Research Questions of the preliminary empirical investigation as part of the domain analysis, my original hypothesis is confirmed: transactional writing is part of the academic domain. In addition, the external experts who were invited to comment on the validity claim of the newly developed exam unanimously stated the proposed transactional writing task fits the academic domain.

The comments of the three international experts helped me decide to include the transactional writing task for Task 1, and redesign Task 2. I revised the specifications and reduced the options for the discursive writing task to one genre, the discussion essay. With these changes, it was possible to keep the writing construct of the Academic test and the General English test the same (as specified by the requirements of profile extension in the *Accreditation Manual*), observe the characteristics of the Academic domain, and make the optional tasks comparable through different test administrations. As for Task 1, to ensure the proficiency level of the task, I specified the number of functions and the number of content points to be used in the task instructions – in addition to observing the academic topic list. Regarding Task 2, in order to observe the needs of students who wish to pursue higher education studies in different academic fields, I specified three academic fields for the optional task: (a) humanities/social science, (b) science, and (c) business/economy. This change made the three optional tasks comparable within the test, and also through the different versions of the test. At the end of Stage 1, I updated the specifications of the writing tasks of the Euroexam Academic test (Appendix 5).

Chapter 7: Completion of Test Specifications, Item Trialling and Pretesting

After the domain analysis in Stage 1, the aim of Stage 2 is to complete the test specifications and the example tasks that will be pretested in Stage 3. The domain analysis presented in Chapter 6 was both narrative and evidence based, I collected student and academic staff interviews and questionnaires for external expert judgement to establish "the knowledge and skills that are valued in the domain" (Perie & Huff, 2016, p. 122). Domain modelling is also evidence focused; I used test taker performance and test taker and rater interviews to see whether the two proposed task types conform to the construct. At this point, it is important to highlight that the verbal protocols and the textual analysis in Stage 2 also serve as the trialling of the exam tasks. When we use test taker and rater feedback, it is only possible to use particular tasks for think aloud and immediate recall, and not the theoretical construct of the genres themselves.

Section 7.5 discusses the activities involved in Stage 3 of the research. They present the results of pretesting of the tasks of the Euroexam Academic C1 level, with special focus on the two tasks of the Writing Paper. On the one hand, I present the statistical analysis of candidate performance data which is used to establish the scoring validity of the Academic test using the original accredited rating scales; on the other hand, I connect the pretest report with the results of Stage 1 and 2 of the development project and point out their mutual relevance. The section on pretesting describes the process and method of data collection, the analysis of the questionnaire responses, and gives a description of the participants. I briefly comment on test taker performance on the Listening and Reading papers of the written part of the test as the objectively marked tasks of these papers allowed me to use them as reference points and compare test taker writing performance to them. Following that, I turn to analysing performance data in relation to the self-assessment questionnaire. At the end of the chapter, the pretest statistics are used to illustrate the argument for reducing subjectivity and rater effect in the assessment of the Writing tasks.

As mentioned above, in trialling and pretesting I used the original accredited rating scales of Euroexam International designed for the assessment of C1 level writing (Appendix 6). For the purposes of Hungarian accreditation, the assessment of writing performances in the trialling phase served a twofold purpose. Firstly, to demonstrate rater behaviour; secondly, to situate Euroexam assessment in relation to the CEFR (Council of

Europe, 2001). Stage 3, pretesting test items normally serves the purpose of standard setting and establishing the scoring validity of the test. Shaw and Weir (2007, p. 6) uses scoring validity for writing tasks as a superordinate term which includes all aspects of reliability, such as the rating criteria of the rating scale, consensual agreement of raters, rater characteristics, the treatment of measurement errors. As it is presented in Chapter 3, in the interest of providing valid interpretation of the results of students' writing production, test providers routinely observe these aspects. By observing the multidimensional, cognitively complex and challenging nature of subjective marking, and aligning the assessment to external standards, test providers aim to reduce the effect of construct irrelevant impacts (Messick, 1994, pp. 14-15) and minimise the level of subjectivity and rater influence in the rating process (Eckes et al., 2016, p. 155). Although scoring validity, as specified by the Educational Authority in Hungary, may be established through trialling and pretesting, thus the results and conclusions of the Stage 2 and Stage 3 fulfil the requirements of Hungarian accreditation, I identified issues with rater behaviour and the use of the rating scales. Due to the sequential nature of the research-based development project, the evidence gathered through these two stages were also used to raise further issues in connection with the validity of the rating process and are built in the design of the following stage (Chapter 8).

7.1 Task characteristics

The tasks of a test paper are designed to map the test taker's proficiency. Therefore, the function of domain modelling is to design tasks based on the information that was gathered in the course of the domain analysis in Stage 1. The tasks are preliminary examples for the new test tasks and are used to pilot the test and item development process (Perie & Huff, 2016, p. 121). Evidence on student proficiency may only be observed through trialling, i.e. using a preliminary task. It is through trialling that we may ensure that test tasks are unambiguous and sufficiently focused on what we want to test (Weir, 2005a, p. 125).

According to Bachman (1990), there are five categories to consider when establishing what kind of task to design. The task characteristics are (a) the testing environment; (b) the test instructions; (c) the nature of the input; (d) the nature of the response expected; and (e) the interaction between the input and the response. As pointed out by Purpura (1999, p. 16), the fifth category of Bachman comprises random factors. Random factors may arise from unexpected events during a test, and the interaction

between task characteristics and the assessment. The semi-structured interviews I decided to carry out with test takers after they completed the task were designed to explore these factors.

The other factor to be observed in the course of domain modelling and trialling is rater behaviour since unbiased and fair assessment is an integral part of test validity. Therefore, it is important to gain insight into what raters do when scoring a writing product, and whether the tasks elicit typical rater behaviour (Lukácsi, 2013, p. 177). To collect information on these two rater-related aspects of trailing, I designed verbal protocols serving as a basis for finalising the specifications for the Euroexam Academic writing test.

7.2 Test taker characteristics: the writing ability of Hungarian students

Although Euroexam International offers language tests in 80 countries worldwide, the majority of the test takers of the Budapest based exam centre is Hungarian. The first language of test takers is an important point to consider even when the test itself is monolingual (i.e. does not contain tasks that are based on mediation or translation). Bárdos (2003, pp. 31-32) draws attention to the interaction between L1 and L2 that needs to be taken into consideration in test development and task design. The target population of the Euroexam EAP test is defined as young adults who are non-native speakers of English and who wish to apply to English medium higher education. The EAP test is planned to be offered in several European countries, but it is expected that Hungarian students would constitute the larger part of the test takers.

As test takers' individual characteristics have been argued to greatly their performance on tests, test designers should take these characteristics in consideration in language test construction (Bachman, 1991, p. 675). Target performance may be elicited and made use of for task design if test taker behaviour and their attitude to the testing situation are observed in the trailing stage.

The question of Hungarian students' writing skills in L2, especially essay writing, has long been a research topic (Bukta, 2007; Fűköh, 2016; Kiszely, 2003; 2006; Molnár 2002; 2009,). Bukta (2007) speaks about "worrying results" (p. 108) of research concerning the students' writing skills development. Molnár (2002) comes to the conclusion that there is no discernible difference in the writing skills of 7th and 11th grade students. Based on her findings, she hypotheses that low achievements in L2 writing is

based on shortcomings in L1 writing instruction (Molnár, 2002, p. 193). Kiszely (2006), when comparing L1 and L2 writing abilities, reveals that Hungarian undergraduates face difficulties in writing in higher education because they have no utilisable knowledge of the different structural units of a text.

The development of writing skills is stated to be an important objective of the *Hungarian National Curriculum* (2012) at each level from grades 1 to 12. However, writing skills as a particular subject is only present in grades 2, 3 and 4. It seems that in the rest of their education, students are supposed to practise writing in literature classes. As a result of this, teachers tend to mark the written assignments based on the content and not the quality of textual organization. Molnár (2009), testing her hypothesis, argues that the curriculum of the Hungarian Language and Literature subject is mainly culture and literature centred, writing skills as such are never assessed; students are taught to focus on the interpretation and aesthetic categories of literary texts as if separable from how students learn to articulate (valid) argumentation.

Hungarian students might be disadvantaged in English language education as regards academic writing skills not only for the shortcomings of the curriculum but for the difference of the essay writing conventions in this (tacit) curriculum. They differ from those of English-speaking countries in significant ways. High school teachers do not usually have a clear idea of the genre conventions in writing, and even state accredited experienced raters have problems with what they reward or penalise in the course of assessment of writing tasks (Lukácsi, 2013). On the basis of these studies, we may conclude that the majority of Hungarian students in higher education have little and inadequate knowledge in the field of writing. As Fairclough (1999) claims, “People practically need to know such things because not knowing them makes it harder for them to manage various parts of their life” (1999, p. 73). This means that Hungarian students at the tertiary level will be unable to perform their studies without the basic skills and practice in the different discourse types which are necessary to conduct their studies at a university.

7.3 Trialling and qualitative data analysis

Domain modelling in Stage 2 is concerned with observing what students who are potential test takers know and connecting the results of the observations to an argument about the test (Mislevy & Riconscente, 2005). In section 7.3.1 and 7.3.2, I move on and present my

empirical data collected in the course of task completion and test taker and rater verbal protocols in order to explore test taker and task characteristics. Domain analysis in Stage 1 focused on the skills and abilities, real world situations, features of these situations, whereas in the context of domain modelling I connect the findings of in Stage 1 to Stage 2 and focus on (a) proficiency, (b) evidence, and (c) the task as outlined by Pearlman (2013, p. 230). Assessing the level of proficiency entails the mapping of test taker abilities, supported by evidence based on observation and the examination whether the proposed task is suitable for the demonstration of the proficiencies. For this, I consulted potential test takers and raters of Euroexam International and collected and analysed qualitative data.

7.3.1 Test taker performance and verbal protocols

The student participants of the trialling phase were undergraduates at Hungarian state universities and students in an International Baccalaureate (IB) Diploma Program (N = 6). The criteria I considered when inviting undergraduate university students (n = 3) required that they have all passed a B2 level English for Specific Purposes test, and they were all planning to continue their studies in an English medium Masters program. The IB students (n = 3), already in an English medium program had B2 level General English certificates and were targeting university studies abroad, which entailed English medium education. The students were informed about the purpose of the research and they all signed a consent form.

In order to interview students about their experience with completion a writing task, domain modelling involves trialling a specific task. The researcher's observations in the course of task completion are required to support the validity claim (Pearlman, 2013, p. 229). The students first were to complete the two tasks: Task 1, *formal transactional email* to a professor and Task 2 *a discussion essay* either in the field of humanities, science, or business in 60 minutes, i.e. the allocated time proposed for the two tasks altogether.

Based on the findings of Stage 1, I designed the two particular tasks of the Writing Paper for trialling. The first one, the compulsory task of writing a formal email to a professor, prescribed two functions to be included, namely explanation and justification, and four content points to refer to, which were (a) reasons for poor class attendance, (b) individual preparation for midterm, (c) issues with essay, and (d) new deadline for essay. The second task, the *discussion essay* comprised three elective essay topics in the fields of (a) humanities: "Participation in a student government greatly benefits your future career.",

(b) science: “It is impossible to improve people’s standard of living without using non-renewable energy resources.”, and (c) business: “The easy money effect of credit cards stimulates overspending”.

In accordance with the exam regulations, they were provided both monolingual and bilingual dictionaries (*Magyar—Angol Kéziszótár*, *Cambridge Advanced Learner’s Dictionary*). During their task completion, I observed their dictionary use, and how much time they devote to each task. After task completion, I conducted one-on-one semi structured interviews with them in Hungarian to explore their own evaluation of their performance and the test task, and the approach they took when completing the tasks. My questions focused on (a) their feelings, (b) their planning strategies (content and structure), and finally (c) their language level (grammar, vocabulary and conjunctions). The preliminary list of questions can be found in Appendix 7. All the interviews were recorded and transcribed. In Section 7.3.1, I provide the respondent texts in my translation. I used MaxQDA for data analysis, which involved identifying and categorising the text content based on the preliminary questions.

Regarding self-evaluation of the test takers, all six students had positive feelings about completing the tasks. Only one student reported feeling a “little uncomfortable” due to being observed during task completion. Other than that, they found task completion “nothing special”. Interestingly enough, five students out of the six were positive about passing the test.

The test takers in the trialling group agreed that they found Task 1, *formal transactional email* much easier because they could rely on their own everyday experiences. Only one test taker reported negative feelings about the first task:

Text: **Test_taker_6 Task 1**
Category: **Feelings**

> …the name of the task ‘transactional writing’ sounded unfamiliar, so I was a bit stressed out whether this would imply a simple email.

As for the content and structure, they pointed out that there was no need for planning in the completion of Task 1, whereas for Task 2, when they had to write from their knowledge and experience, they all reported that they tried to think over what they wanted to write. One student, Test taker 1 even reported writing an outline for the task similarly to Task 1, where the required functions and content points were given.

Text: **Test taker_1**
Category: **Planning**

I knew I wouldn't have time to draft the whole text, but I jotted down a few points as an outline.

Text: **Test taker_2**
Category: **Planning**

The task was quite specific about what to include, so here I only focused on how to link the given content points.

Another aspect the students mentioned in connection with content was the formality of Task 1. They all found the topic of the email very formal and paid attention to being very polite. This finding cross validates the finding of the preliminary empirical study within the domain analysis about the perceived formality of a transactional email in the domain of higher education. The student respondents in the study pointed out that formality is an essential feature of the emails they write to university members of staff and teachers. Although it is becoming largely examined whether there is a shift towards being more informal in written genres (Fairclough, 2001; Leedham, 2015), it is not evident that greater informality characterises all areas of academic writing (Hyland & Jiang, 2017). Based on my empirical findings in Stage 1 and the relevant literature, I concluded that the level of formality is an essential element of Task 1, *transactional writing*.

As for planning in connection with Task 2, all three IB students reported planning a 5-paragraph essay in their heads. They pointed out that this is the format the IB courses require, so writing an essay was not challenging for them. At the same time, the BA students did not plan the text ahead, nor did they write an outline of the essay. This observation conforms to the finding of the literature cited above in connection with the shortcomings of Hungarian primary and secondary education, and the results of Kiszely's (2003; 2006) research in connection with the writing ability of university undergraduates.

As regards test takers' language level, they all reported paying special attention to complex grammatical structures, especially the use of conjunctions.

Text: **Test taker_3**
Category: **Language**

> My experience from the B2 exam prep courses is that if you want to pass a language test, you have to produce a text packed with language at level. I think I succeeded in that.

Their feelings about the range of vocabulary they used in their answers were positive; the use of the word "sufficient" was salient.

Text: **Test taker_5**
Category: **Language**

> I'm not worried about my vocabulary; I think it is sufficient for level C1.

It appears that the students felt positive about their performance, they did not feel stress during task completion, and they thought that they are at the required level as regards language and task fulfilment.

The students were also asked questions to evaluate the test taking experience, comment on the tasks and how they found completing them. None of the students opted for the science topic, the three university students picked the finance, whereas all IB students chose the humanities one.

As I observed no dictionary use with any of the students, I asked them in the interview about the reasons. Although they managed to finish the tasks in the allocated time of 60 minutes, all participants pointed out that one of the reasons for not using a dictionary was the pressure of time. The other reason is in connection with their self-evaluation: they all thought their active vocabulary was sufficient for task completion.

The typically emerging observation in connection with Task 1 was that they found it very strange to write an email on paper and, more importantly, that they found the task instructions insufficient. As for Task 2, they found the task instructions clear and they were happy that they could select the topic for themselves from the three field specific options.

Text: **Test taker_2**
Category: **Task characteristics – Task 1**

> It would have been clearer if I could use an email template.

Text: **Test taker_1**
Category: **Task characteristics – Task 1**

> The name of the professor wasn't given, I came up with a random name, but I wasn't sure what I was supposed to do.

Text: **Test taker_6**

Category: **Task characteristics – Task 2**

I read all three options, and considered my background knowledge.

Text: **Test taker_5**
Category: **Task characteristics – Task 2**

…it was clear that the arguments for and against have to be balanced.

In general, the students had positive impressions of the test tasks. They all claimed to have practised tasks like these either as course content, or real-life tasks. The student comments regarding the task instructions and layout were taken into consideration in the finalisation of the test specifications.

In addition to the verbal protocols, the student writing products were assessed in two different ways. First, I carried out a textual analysis of student writing products, which was followed by an overall judgement based on holistic impression of each script. This overall judgement served as the basis for the selection of two scripts that were assessed through think-aloud protocols by accredited raters of Euroexam using the writing scales. The purpose of the textual analysis was to reveal whether the self-assessment of the students conformed to their achievement. My assessment method used test taker self-evaluation comment categories and an objective list of textual features, so-called text indices, regardless of the score categories of the C1 level Euroexam writing scale (Appendix 6). The use of automated text analysis tools could add invaluable insight to test taker performance, as the algorithms they use produce textual analysis on various language and discourse levels (Graessler et al. 2011, p. 34). The linguistic features of large sample data may make the quality of writing products comparable across test versions with the use of Coh-Metrix or Compleat Lex Tutor (Crossley & McNamara, 2011; Taylor, 2010). This time, however, the text indices were coded in MaxQDA using a small sample to combine qualitative and quantitative data. The use of detailed categories made human annotation more suitable for the purposes of the present research. The indices I introduced were based on studies which focused on detailed textual analysis of student writing (Endres, 2012; Varner et al., 2013).

On average, the number of words for Task 1 was 207.6 (*SD* = 29.79), whereas the answers for Task 2, as expected, were significantly longer, with an average of 262.4 words (*SD* = 27.07). In my analysis, I set up the following categories for the purpose of examining text complexity. Category (a), the number of paragraphs, more specifically, the

number of visible paragraphs in both tasks; Category (b), the content elements, which in the case of Task 1 was based on the given content points and functions (four altogether), while for Task 2, it meant the number of content elements as specified by the task rubric. The third, Category (c), included reference devices based on three different types of references: pronominal (*which, it*), demonstrative (*that, this*), comparative (*other, more*). I observed conjunctions (*so, after all, furthermore*) as Category (d), and finally in Category (f) I looked at grammatical devices (*indeed, having said that*) In Category (e), and lexical devices, i.e. outstanding vocabulary representing C1 level (*vain effort, attendance*). I checked the CEFR level of the outstanding lexical items with the help of *English Vocabulary Profile* (English Profile, 2015). Using the above categories, I counted their occurrence in each of the student writing products (Test taker 1-6) broken down into Task 1 and Task 2. The student scores for each task and category are displayed in Table 6.

Table 6
Textual Features of Student Writing Products

Categories:	**(a)**		**(b)**		**(c)**		**(d)**		**(e)**		**(f)**	
	Number of paragraphs		**Content elements**		**Reference devices**		**Conjunctions**		**Grammatical devices at level**		**Lexical devices at level**	
Task number	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Test taker_1	1	4	4	4	16	25	7	8	7	19	5	8
Test taker_2	6	3	4	2	14	19	6	5	0	9	2	0
Test taker_3	5	3	4	2	8	16	9	10	0	4	0	4
Test taker_4	3	1	3	1	7	9	4	5	0	2	0	5
Test taker_5	2	5	4	5	13	22	4	6	5	8	6	7
Test taker_6	4	5	4	5	17	21	6	7	7	12	4	8
M	**3.5**	**3.5**	**3.8**	**3.2**	**12.5**	**18.7**	**6.0**	**6.8**	**3.2**	**9**	**2.8**	**5.3**
SD	**1.9**	**1.5**	**0.4**	**1.7**	**4.1**	**5.6**	**1.9**	**1.9**	**3.5**	**6**	**2.6**	**3.1**

Studying the scores of the individual students in the six categories of textual features, it becomes clear that their writing performance shows diversity. Comparing students' self-evaluation with the standard deviations calculated from the average scores reveal that their

concepts of quality writing are the closest in terms of paragraphing (a), content elements (b) and the use of conjunctions (d). It seems, however, that they must have very different understanding of advanced use of reference devices (c), C1 level grammar (e) and lexis (f) as there is a discrepancy between the scores and their self-satisfaction with their performance they reported in the interview.

Based on the scores that altogether should define text qualities, I assigned overall judgement scores based on holistic impression from 1-3 to each individual's writing performance, where 1 = weak, 2 = satisfactory, and 3 = outstanding performance (Table 7).

Table 7
Overall Judgement of Student Performance

	Overall judgement	
Task number	**1**	**2**
Test taker_1	3	3
Test taker_2	2	2
Test taker_3	2	1
Test taker_4	1	1
Test taker_5	2	3
Test taker_6	2	2

The overall judgement served as the basis for choosing the two scripts for assessment of rater think-aloud. I decided to use two very different student scripts for scale-based assessment. I chose Test taker 1 and Test taker 3, a high and a low performer respectively in order to see what raters reward and penalise and also to avoid the assessment practice of central tendency. My choice fell on these two test takers for another reason: they both opted for the topic (c), in the field of business.

7.3.2 Euroexam rater verbal protocols

Assessment of rater behaviour with the help of a rating scale has been a well-researched topic in the past decade. Benke (2007) in her dissertation analysed rater comments in the course of scoring and identified the categories of the rating scale that are challenging for raters. Bukta (2007) examined rating processes of EFL compositions of teacher trainees

and analysed rater behaviour. Based on empirical data, she reports the existence of an imbalance and rarity of comments on test organisation. Lukácsi (2013) researched Euroexam raters to find out about their understanding of coherence and cohesion. His conclusion, based on the comparison of four markers' comments, is that all four of them defined coherence and cohesion in a different way, the only common element was 'cohesive devices'. As for paragraphing, he found that it "remains largely undefined but by implication [it is reduced] to layout" (Lukácsi 2013, p. 207).Varner et al. (2013) conducted empirical research with the involvement of 126 students who were asked to write timed impromptu Scholastic Aptitude Test (SAT) style essays. Having finished the task, students had to evaluate themselves on a scale of 1 to 6, where 1 was the weakest and 6 was the strongest point. After scoring the essays, they found that there was a significant evaluative misalignment between students and raters. Their study shed light on the different perceptions students and raters have about the quality of writing.

The analysis of rater think-aloud protocols in the present research is a less detailed one, the focus of my analysis is on how raters approach the academic tasks, the differences among raters' perceptions, and the misalignment between raters and students regarding their understanding of the quality of the writing products.

Three experienced accredited raters of Euroexam International took part in the research. They were briefly informed about the construct of the new Academic test and the purpose of the present research. As mentioned above, based on my overall judgement, I choose two scripts of very different writing quality (Test taker 1 and Test taker 3).The three raters were given the same two test taker scripts and were asked to assess them using the accredited C1 level writing scale of Euroexam International (Appendix 6) using a think-aloud technique in Hungarian. The sessions were individually recorded and took place in March 2019. I present the examples from rater responses in my translation.

The rating time in the think-aloud protocols varied between 15-20 minutes per text. Judging by Norton's (1990) study, the average time to mark a 200-250-word texts would be around 4-5 minutes, which makes the amount of time spent on rating the two tasks 10 minutes. This calculation is in line with the fact that raters of live administrations spend one hour with marking a maximum of 5-6 Writing Papers. Benke (2007), however, found that verbalisation of the rating process may result in significant differences between the rating time of different raters (p. 64).

The raters first verbalised their impressions of the texts, then they needed some time for reading, followed by a systematic use of the categories of the rating scale illustrated by examples they read out loud. Although they used the Euroexam rating scale, (Appendix 6), in the analysis of the verbal protocols, I decided to use the same six categories (a-f) I established for the textual analysis of student scripts in Table 6 for the analysis of the three raters' verbal protocols, so I coded them in MaxQDA accordingly. The reason for this decision was twofold. First, I wanted to use categories that are comparable; second, the writing scale descriptors used by the Euroexam are too vague, so I found it more beneficial to go along with the more explicit categories of textual assessment I introduced in Section 7.3.1. In addition to this, I recorded the scores they awarded based on the Euroexam C1 level rating scale in Table 8. (The detailed analysis of the Euroexam C1 level writing scale can be found in Chapter 8.) Based on the verbal reports, I complemented the categories with (g) initial impression as a seventh one, since all raters started the think aloud protocols by looking at the text as a whole and judging by the handwriting. As the rating scale of the two tasks is the same, and the raters followed the same routine when rating the two tasks, I analysed the verbal reports of the rating process of the Task 1 and Task 2 together.

As regards the tendencies inferred from the protocols, I could see that the initial impression voiced by the three raters was mainly concerned with legibility and the number of visible paragraphs.

> **Text: Rater_2**
> **Code: Initial impression – Task 1**
>
> This one is easy to read, there are no corrections or illegible scribbling in it.
>
> **Text: Rater_1**
> **Code: Number of paragraphs – Task 1**
>
> I can see that the test taker knows how to paragraph a text.

Regarding the category of content elements of the assessment, it was only an issue for Task 2, the discussion essay. As introduced above, I made sure at the end of Stage 1 that the specifications prescribe fours content points to be given based on which the formal transactional email can be composed. As opposed to this, the discursive writing task is introduced by a short, one-sentence prompt with no requirements of specific content

elements. As a consequence, the rater verbal reports concerning the category of content elements was very different for the two tasks.

In case of Task 1, they were looking for the presence of the given content points, and two of them (Rater 1 and Rater 3) were ticking the ones they found on the test paper. As opposed to this common understanding, the raters had very different ideas about the content of the discussion essays. It was very typical that they pointed out that the rating scale is the same as for the C1 level Euroexam General English Test but kept reminding themselves from time to time that they are dealing with Academic Test tasks, which affected the harshness of their rating.

Text: Rater_1
Code: Content elements – Task 2, Test taker 1

The conclusion is pretty good, but now that I think of it… this is supposed to be Academic, then… maybe not so good.

Furthermore, the raters were not sure what level of personality the genre of the discussion essay allows, or what content points they can expect.

Text: Rater_1
Code: Content elements – Task 2, Test taker 1

This is a good one, but it contains only general remarks: "*there are people*", "*in general*". I would prefer something more concrete.

Text: Rater_2
Code: Content elements – Task 2, Test taker 1

I'm not sure, is this academic enough? The writer did not elaborate on many points.

Text: Rater_3
Code: Content elements – Task 2, Test taker 3

…it's too personal; I don't think it's appropriate to write about personal experience.

Based on the above comments, it became evident that the raters had very different ideas about what content elements they were looking for, whether a personal or an impersonal essay is better, or what constitutes the academic nature of a text.

I also noticed that the surface level of visibility and the content level for paragraphs (Category (a)) often merged in the course of the evaluation protocol. Despite the fact, for

instance, that the paragraphs were not visible in one of the texts, the raters forced themselves to find structural elements based on the content.

Text: Rater_3
Code: Number of paragraphs – Task 2, Test taker 1

This test taker did not use paragraphs, but now that I read it in detail, I can see that they are indented, so after all the structure is good.

Text: Rater_2
Code: Number of paragraphs – Task 2, Test taker 1

There are no paragraphs visible but as I read it, it is coherent.

Considering that the number of paragraphs is an objective category, it was surprising that the three raters approached it in very different ways. The underlying reason for this might be one of the shortcomings of the rating scale, i.e. the wording of the descriptors use vague categories which are difficult to relate to.

As to the categories of reference devices (c) and conjunctions (d), the rater comments were even more varied. They noticed the use of reference devices and conjunctions for the rating scale category of Cohesion, but they related to these very differently.

Text: Rater_2
Code: Conjunctions – Task 2, Test taker 1

There are conjunctions, but I don't think they are at level. Like *moreover*, it is very simple.

Text: Rater_1
Code: Conjunctions – Task 2, Test taker 1

This one uses linking words, I like *moreover* and *however* so Cohesion is 4.

Text: Rater_2
Code: Conjunctions – Task 1, Test taker 1

OK, they use *firstly*, *secondly*, but these are just panels they are pre-prepared for.

Text: Rater_3
Code: Conjunctions – Task 1, Test taker 3

> I can see linkers in the text, but the text itself is so simple, and there are too many grammatical mistakes, so I'm not going to award a high mark on Cohesion.

The above four examples, which refer to the use of the conjunctions highlight the controversies of rating with the Euroexam rating scale. Despite mentioning the same conjunction (*moreover*) the perception of their use and quality is very different for Rater 1 and Rater 2: the former marked down the test taker for the use of simple conjunctions like *moreover*, whereas the latter was satisfied with *moreover* and *however*, and awarded 4 points out of five. Rater 3 acknowledged the use of *firstly* and *secondly*, but put forward that these are used routinely and did not award a high mark for Cohesion. Furthermore, Rater 3 is penalising the student regarding the Category of Cohesion based on the grammatical mistakes.

The two last textual categories in Table 6, (e) grammatical devices and (f) lexical devices, were almost identical with the categories of the Euroexam C1 level rating scale (Grammatical Range and Accuracy and Lexical Range and Accuracy), and these were the ones where they had very similar perceptions. The raters were consistent in pointing out the grammatical mistakes and the lexical shortcomings of the texts.

> **Text: Rater_1**
> **Code: Grammatical devices – Task 1**
>
> The grammatical structures are very simple in this one. I would say this is a B2 text.
>
> **Text: Rater_1**
> **Code: Grammatical devices – Task 2**
>
> The vocabulary of the text is rather simple, but I can see some good examples, like *repercussion*, or *vain effort*.
>
> **Text: Rater_2**
> **Code: Lexical devices – Task 2**
>
> Oh, wow, the candidate used the word *repercussion*, it deserves a 5 for vocab.
>
> **Text: Rater_3**
> **Code: Lexical devices – Task 1**
>
> [reading aloud] *I do not want to say a reason*… this sounds inaccurate.

As mentioned above, the scores the three raters awarded for Task 1 and Task 2, using the accredited C1 level rating scale (Appendix 6) are displayed in Table 8. The criteria of the rating scale are as follows: (a) Task Achievement, (b) Appropriacy, (c) Coherence, (d) Cohesion, (e) Grammatical Range and Accuracy, and (f) Lexical Range and Accuracy. When we look at the scores for Student 2 and Student 3, it is clearly discernible that rater agreement is lower for the high performer student (Task 1 *M* = 19; *SD* = 5.5, Task 2 *M* = 19; *SD* = 4.6.), whereas the raters tended to agree more in the scores for the low performer student (Task 1 *M* = 7.7; *SD* =.6; Student 2 Task 2 *M* = 7.3; *SD* = 1.5).

Table 8
Individual Rater Scores

	Student_1												**Student_3**											
	Task 1						**Task 2**						**Task 1**						**Task 2**					
Rating scale Criteria	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f
Rater_1	4	5	3	4	4	5	2	3	3	3	2	3	2	1	1	1	1	1	2	2	1	2	1	1
Rater_2	3	2	2	2	2	3	4	3	2	2	2	3	2	2	1	1	1	1	2	1	1	1	1	1
Rater_3	3	3	3	3	3	4	4	5	4	3	4	4	2	2	1	1	1	1	1	1	1	1	1	1

The individual rater scores for Student 3 range from 20% to 26.67%, which means that the raters largely agreed that the performance of the student was unsatisfactory, nevertheless these score ranges never appear in live administrations (see Chapter 8 Figures 12-14).

My findings for C1 texts conform to those of Lukácsi (2013) for B2 texts: he found that Euroexam raters of B2 texts "varied in their scoring behaviour, their construct interpretations, and their severity in the think-aloud session" (p. 232). The behaviour of raters of C1 texts in my sample is similar. When the three raters are referring to the rating scale, they have very different understanding of the criteria. Although all raters identify the textual and language elements on the basis of which scores can be awarded, they are unsure about how to weight them in their assessment. The vague descriptors of the rating scale hinder the objectivity of the rating process and thus the reliability of the test scores.

7.4 Finalising the specification and the test items

The main aim of Stage 2 of the development process was to complete the specifications using the empirical data that was gathered through domain modelling and trialling. Based on the qualitative data, we may conclude that we had ample evidence to "move from claims to test specifications" (Perie & Huff, 2016, p. 128).

As a result of Stage 2, I redesigned Task 1 based on the students' comments. I specified the name of the recipient of the email and added an email template to make the task instructions unambiguous, resembling the electronic layout. Furthermore, based on the student interviews, the context validity of the two writing tasks could adequately be established. The students who were identified as potential test takers put forward that they were familiar with the task types and did not have any difficulties with task completion. Furthermore, the university undergraduates confirmed that they have hands-on experience with fulfilling similar tasks in the academic domain. I also used the student comments and suggestions to complement the task specifications, which will help the future item writing process. As a result, a detailed task specification was completed for the Euroexam Academic Test together with the two sample writing tasks (Appendix 8). In Stage 3, large-scale pretesting, I used the format and layout of the tasks as they appear in Appendix 8.

7.5 Pretesing and quantitative data analysis

Stage 3 of the validation process involved large scale data collection and evidence-based analysis of test taker performance. The aim of this stage was to check that test tasks work as intended so that the standard level (C1) of the test could be set, the relationship to the CEFR could be established and the validity of the test could be demonstrated (Council of Europe, 2009). To ensure all this, I used a sample size that allows statistical data analysis using Classical Test Theory (CTT). In addition to large sample size and modelling the testing environment, I also payed attention to administer the pretest with a population that matches the target audience of the C1 level Euroexam Academic test.

7.5.1 Methods and data collection

The pretest and the questionnaire data were collected in five different Hungarian high schools using convenience sampling. The target population was defined as young people or adults (16+) who wish to apply to study on an undergraduate/graduate programme where the language of tuition is English. Five schools in and outside Budapest were contacted

where students usually reach the level of C1 by grades 11 and 12 – based on teacher reports. The aim was to contact schools outside the scope of the "traditional Budapest elite education" – as reflected in the rank of schools published annually. The data collection took place in May-June 2019. I asked the language teachers in the schools to administer the Euroexam Academic test, a paper-based test. The test takers used answer sheets to enter their answers, which were computed to excel spreadsheets using IBM SPSS 24 in July-August 2019.

The test takers of the pretest were also asked to fill in a three-part questionnaire (Appendix 9) in Hungarian after the completion of the exam so that I could have access to the respondents' thoughts and feelings in connection with the Academic test. The first part of the questionnaire asked for personal data, contained open ended questions about students' language learning background and language exam certificates. The second part also used open ended questions, inviting the test takers to evaluate their own pretest performance. (The results of their self-evaluation are analysed with performing a chi square test.) The third part consisted of four Likert scale items (1-5) in relation to the content and form and content of the test. (The results are provided in Table 12 and Table 13)

The aim of the pretest was to find a sample population of students who are at level B2+/C1, have already passed general B2 or C1 exams, and who planned to continue their studies in English language higher education. I wanted to secure some 100-150 students taking the test for all the Reading, Listening, Writing and Speaking papers. They eventually turned out to be 136, representing five schools. Although my initial aim was to go beyond the scope of the top schools of the high school rank, this criterion was only partly fulfilled, as one of the five high schools (ELTE Radnóti Miklós Gyakorló Általános Iskola és Gimnázium) was proposed to be third in the list of best Budapest high schools in 2018 (Eduline, 2017). The pretest population shows the following distribution:

24 students from Magyar—Angol Tannyelvű Gimnázium, Balatonalmádi

15 students from Deák Téri Evangélikus Gimnázium, Budapest

34 students from Kazinczy Ferenc Gimnázium és Kollégium, Győr

21 students from Madách Imre Gimnázium, Budapest

42 students from ELTE Radnóti Miklós Gyak. Ált. Isk. és Gimnázium, Budapest

The gender distribution was balanced, altogether 67 female and 69 male students participated in the pretest. The age range was between 17 and 21 years of age (M =19.29; SD = .92), with an average of 8 years of English learning background. 63% of the students reported holding a language exam certificate ranging from B2 to C2 (57 B2, 26 C1 and 3 C2). The raters of the pretest papers were 4 experienced and trained raters of Euroexam International.

7.5.2 Discussion of test papers and results

The 'written papers', i.e. the parts of the test which are completed in writing, of the Euroexam Academic Test consist of Listening, Reading and Writing. At the end of Stage 2, I finalised the specifications and the item writer documentation for the writing tasks, while the other members of the Academic Development Team did the same for the Listening and the Reading papers of the Academic Test. The items of the three papers were written and compiled using these documents. The focus of the dissertation is the development and validation of the writing tasks, but the procedure of setting the standard of the new test as it appears in Stage 3 requires the comparison of test taker performance regarding each of the written papers of the test, all the more so because the Listening and the Reading papers contain objectively marked tasks, whereas the Writing paper contains subjectively marked ones. Furthermore, the Listening and Reading papers always contain common items, in other words repeated tasks so that the two samples are statistically equivalent (Council of Europe, 2009, p. 85). This measure ensures that the different tests batteries are comparable across administrations.

The task types for the Writing paper are the same for the General and Academic C1 tests (transactional writing and discursive writing), however, the Euroexam Academic Writing paper tests the writing skills of students within an academic domain with a formal transactional email in an academic context and a discussion essay tasks in three different fields of study. The tasks of the Listening and the Reading papers of the Academic test slightly differ from those of the general C1 Euroexam. One task of the Listening paper (Task 2) and two tasks of the Reading paper (Task 2 and Task 3) were the ones that were specifically developed to test Academic skills by the other members of the Academic Development Team, the remaining tasks were repeated from previous administrations, the topics of these tasks were exclusively chosen based on the list of topics in the *Guide for Item Writers* (Euroexam International, 2018c).

As mentioned above, I used the results of test takers on the Listening and Reading papers and the correlation between the repeated tasks and the academic tasks to cross validate the results of my analysis of the subjectively marked Writing tasks. For this reason, I report on how the test takers performed on the other two papers as well. The basis of the comparison was the internal document titled *Euroexam Academic Pretest Report: Reading and Listening Papers* (Lukácsi, 2019a). Using the statistical data of the report allowed me to draw conclusions and make comparisons with the statistical analysis I carried out in relation to the Writing papers.

The Listening Paper had 25 items in three tasks. The first task was repeated from September 2014 and the third task was repeated from October 2016. The second task was a unique task specifically designed for the Academic exam. The reliability of the test paper is acceptable (Cronbach's α = .763). The pass marks in Table 9 show a discernible difference for the three tasks.

Table 9
Pass Marks for Listening Tasks

Repeated Task 01	69.75
Academic Task 02	27.25
Repeated Task 03	77.54

The repeated tasks seem to have been easy for the population, whereas the Academic task was extremely difficult.

The Reading Paper consisted of 20 items in three tasks. The first task was repeated from May 2014. The reliability of the paper fell short of what is expected (Cronbach's α = .697). The relationship between the Academic task and the repeated task shows a similar pattern to the results of the Listening paper (Table 10).

Table 10
Pass Marks for Reading Tasks

Repeated Task 01	77.63
Academic Task 02	68.40
Academic Task 03	45.19

These values are comparable to the ones reported about the Listening paper. Although the targeted CEFR level of the test is the same, the tasks which are specifically designed for the Academic Test proved to be more challenging for the population. At the same time, the

correlation between the repeated tasks and the Academic tasks shed light on a discernible difference between the population of the repeated tasks and the Academic Test pretest population. The repeated tasks of the Listening and Reading papers worked very similarly to the live administration; however the mean percentage of test takers who passed the live administration was 69% as opposed to 81% in case of the Academic pretest population, which shows that the "pretest population was on average 12% more able than the live administration" (Lukácsi, 2019a, p. 5).

As for the subjectively assessed Writing papers, the difficulty was similar to live administrations. The correlation between the first and the second rater was relatively low (r = .754). Inter-rater agreement α_K = .739. Both these values fall short of what is expected or acceptable (see Chapter 5). Table 11 shows the mean, the standard and deviation the pass rate for the 3 test papers.

Table 11
Means of the Test Papers

	Listening	**Reading**	**Writing**	**Result**
M	59.16	68.34	63.30	63.64
SD	16.57	16.34	15.56	12.92
pass rate	50.73	69.85	63.23	59.55

Although the pass rates fall in the acceptable range, and the average pass rate is 59.55, which conforms to the cut-off score of 60%, it is rather unusual that the writing results are the closest to the cut-off score, and show no difference between the General C1 and the Academic population, which might be merely an indication of rater behaviour (Lukácsi-Fűköh, 2018; 2019).

7.5.3 Test taker opinion

As for the test takers' opinion of the test and their performance, it is the second and third part of the questionnaire that are of relevance for my research. The open-ended questions of these two parts aimed to explore the test takers' perception of the qualities of the test behind easiness and difficulty and also what test takers think about the form and the content of the test. The second part aimed to reveal whether the test takers are able to realistically predict their own performance and at the same time rate the difficulty of the test papers. The test takers' perception of the difficulty of the test papers shows a very similar pattern to the results: 75% thought that the Listening paper was the most difficult,

whereas only 9% thought that it was the easiest part. To compare test taker perception of the facility of the tasks to performance data, a chi square test of independence was performed. We can see that although there is strong significant relationship between the test takers' predictions concerning their own general performance and their results ($\chi 2$ (1), N = 136) = 8.28; p = .003), this perception is completely different for the writing tasks. The results of the test show that there is no significant association between students' perception and the results of their own writing performance ($\chi 2$ (1), N = 136) = .28, p = .889).

The third part of the Questionnaire asked the test takers to evaluate the form and content of the test regarding four categories on a 5-point Likert scale, where 1 = unsatisfactory, 2 = satisfactory, 3 = average, 4 = good, and 5 = excellent. The averages of the 5-point Likert items are displayed in Table 12.

Table 12
Evaluation of Form and Content – Likert-Scale Averages

Appearance	4.08
Instructions	4.22
Task type	4.36
Content	3.47

The averages show that the test takers were almost equally satisfied with the test papers' appearance and the instructions. The average points awarded for the content of the tasks although lower can still be considered satisfactory as the essay topics of Task 2 of the Writing paper are elective. The answers of the open ended questions regarding the features of the different test papers support my judgement (in Table 13).

Table 13
Salient Reasons for Task Easiness and Difficulty

Listening	easy:	short questions, dictionary use
	difficult:	long; fast; difficult topics, vocabulary, accent; listen to only once
Reading	easy:	straightforward questions; MCQs; well-structured texts; ample time
	difficult:	difficult vocabulary, topics
Writing	easy:	good topics; dictionary use; gives room for creativity
	difficult:	some difficult essay topics; allocated time not enough

It is discernible that the qualities attributed to easiness are in connection with the form, whereas the qualities attributed to difficulty are in connection with the content. I am going

to focus only on the writing component, the domain of my research interest. The formal features that make the test papers difficult (allocated time for the writing) was considered to be changed after the pretest. The suggested timeframe has been changed to 20-40 instead of the original 30-30 minutes. The management of Euroexam International believed that lengthening the allocated time would not result in better quality writing performances. As referred to above, the issue of topic difficulty may be resolved by the fact that the three essays of Task 2 or the Writing paper are elective, and test takers may choose the one that is the most appropriate for them.

7.6 Conclusion

The small-scale studies using verbal protocols in Stage 2 convinced me that the two independent tasks and the task type of transactional writing are suitable for the EAP test. The students in the course of the interviews confirmed that they are familiar with the task types, and practice these during their studies. With highlighting that they write formal transactional emails in the context of their studies, they cross validated the preliminary investigation and the expert judgement of Stage 1.

The results of Stage 2 and Stage 3 provided empirical evidence that the Euroexam Academic test, as regards its form and content together with its relation to the existing accredited C1 level general test and the CEFR, demonstrated a valuable addition to Euroexam International's exam portfolio. The Writing tasks proved to be valid measures of target language level and the academic domain.

Another important outcome of Stage 2 and Stage 3 together, however, raises some concerns about the scoring validity of the two tasks of the Writing paper. The most important requirement of establishing the scoring validity of language test is to make sure that the tasks measure what is defined in the construct. In case of writing tasks, valid measurement heavily relies on the assessment tool and the assessment procedure of writing products. The writing results of the pretest seemed to justify the findings of the trialling phase, i.e. there is a considerable rater effect in the assessment of the writing tasks, which has further implications as regards scoring validity. The rater verbal protocols and the pretest results of the writing tasks revealed that the scoring validity could be improved by designing an assessment tool that leaves fewer opportunities for subjectivity.

The Academic Test being a high-stakes test that affects the future of test takers, it is of paramount importance to ensure that subjectivity is reduced in the assessment of writing

tasks. The use of an objective rating tool is expected to reduce differences among raters. Apart from this immediate result among raters, there is a predicted positive washback effect that will develop students' genre awareness and writing skills and also increase the probability of the correct perception of their writing results.

Based on theoretical considerations and empirical evidence, the need for a more objective rating tool seems to have been grounded. Since the development project is iterative in nature, the issues raised in connection with scale-based rating will be further investigated in Chapter 8 by qualitative and quantitative methods. The detailed analysis of the Euroexam scale-based rating tool, and the checklist development project is presented in Chapter 8.

Chapter 8: Establishing the Scoring Validity of Checklist-Based Rating

Both the requirements of the Hungarian accreditation system and the international recognition standards demand that language tests match internationally recognised proficiency frameworks. The alignment of a test to a common framework, like the CEFR (Council of Europe, 2001) ensures the accountability, and the reliability of the measuring tools. Euroexam considers it vitally important to measure the language knowledge of their candidates in a valid, reliable and non-biased way. Therefore, Euroexam International has adopted results reporting that uses a one dimensional IRT model for equation (Verhelst, Glas & Verstralen, 1995) so that they can guarantee that candidates pass a test in any exam period at the same level of knowledge and skills. As regards the assessment of writing tasks, total scores are calculated from the average of two independent ratings, which may minimise rater discrepancy (Weigle, 2002).

In Stage 2 and Stage 3, the validation process of the Academic writing tasks, the rater think aloud protocols and the results of the pretest shed light on raters' differences across the concepts in connection with the genre and the qualities of the discussion essay within the academic domain. Although rater training and re-standardisation may compensate for rater harshness and leniency, it is impossible to model differences in rating which originate from the beliefs or the cultural and social aspects of particular raters (Weigle, 2002, pp. 70-72). Consequently, based on the rater interviews, I decided to design a level and genre specific checklist-based rating tool with in order to strengthen the awareness of good quality essays, which, in turn, could reduce the time of compulsory training and retraining at Euroexam International, Budapest.

The aim of Stage 4 is to increase the scoring validity and the reliability of the assessment of the essay task and design a tool for assessment that compensates for individual rater characteristics and rater effect. Reliability may be increased by using a tool that is able to distinguish between low achievers and high achievers (Bachman, 1990), while the subjectivity of rating may be taken away by designing a checklist-based tool with dichotomous items, as suggested by the CEFR (Council of Europe, 2001, p. 189). In this chapter I focus on the development process of the checklist-based assessment tool for discussion essays. In the name of developing a non-biased, objective rating tool, I discuss the steps of the development process from document analysis, through teacher task completion and interviews, to large sample testing.

8.1 The need for a non-biased objective rating tool for Euroexam

Reviewing the relevant assessment literature, it becomes clear that the assessment tool of writing ability is almost exclusively the rating scale (Harsch & Martin, 2013). In their theoretical framework, Weir (2005a) and Shaw and Weir (2007) only discuss the advantages and shortcomings of the different types of rating scales, but the use of a checklist is not part of their proposal. Further to this, the shortcoming of rating scales is a prevailing topic in assessment literature, and we may find various suggestions to counteract these (Bachman & Palmer, 1996; Eckes, 2009; Hamp-Lyons, 1990; Harsch & Martin, 2012; Lukácsi, 2018; 2020; Wiggelsworth, 1993).

It has long been agreed that the wording of rating scales often results in individualistic, subjective assessment due to the use of vague language and indefinable criteria (Weigle, 2002), which conforms to the findings of Stage 2 of the present validation research. When listing the advantages and disadvantages of analytic rating scales, Weigle stresses the necessity of "well-articulated levels" (p. 119) within the scale. In Chapter 3, I presented general problems with scales, their use, and the raters; in Chapter 7, I collected empirical evidence for the use of the C1 level Euroexam rating scale. In the current chapter, then, I focus on the detailed analysis particular C1 level writing scale of Euroexam International and the local raters.

The Euroexam rating scale formulates two criteria regarding reliability and validity of rating writing tasks. Firstly, observing the Hungarian accreditation requirements (Government Decree 137/2008 (v. 16)), the rating scale is level specific. The different bands of the scale do not represent different levels but focus on one specific level (B1, B2 or C1) and the scale of 0 to 5 represents the quality of the writing product within that level. Secondly, the bands in the rating scales (Euroexam International, 2020) are undefined: the even score values (2 and 4) on the scales are empty, leaving the raters without a particular descriptor to observe. There are two other features of the C1 level rating scale to point out, the structure and the wording of the descriptors (See the full scale in Appendix 6). I have selected one particular criterion for Task Achievement to demonstrate the issues with the descriptors of the different bands (Table 14). Five criteria of the Euroexam rating scale are directly related to the communicative language activities of the CEFR: (a) appropriacy, (b) coherence, (c) cohesion, (d) grammatical range and accuracy, and (e) lexical range and accuracy. The choice of the Task achievement criterion is motivated for two reasons. One the one hand, unlike all the other criteria, the formulation of this one does not borrow

statements from the CEFR illustrative scales, it is based on an entirely internal development. Due to the communicative approach of the framework, "achieving a task" per se is not part of foreign language skills, so the wording of the descriptors does not come from the CEFR. On the other hand, as the rater verbal protocol has shown, the criterion of task achievement is closely related to the raters' awareness of the genre, a main objective to demonstrate and improve in the present research.

Table 14
Euroexam C1 rating scale for C1 writing – excerpt (Euroexam Level C1, 2020)

5	**Task achieved at a high level** **Intention:** Entirely clear **Instructions:** Completely followed **Effect:** A positive effect on the target reader **Outcome:** Sure to achieve a successful outcome **Content:** All relevant details included. Some original ideas or presentation
4	
3	**Task achieved, some gaps** **Intention:** Clear in most areas **Instructions:** All important ones followed **Effect:** A generally positive effect on the reader. **Outcome:** Likely to achieve a successful outcome **Content:** Many relevant details included
2	
1	**Task unachieved** **Intention:** Very unclear. **Instructions:** Many not followed **Effect:** Negative **Outcome:** Will not achieve a successful outcome **Content:** Omission, irrelevance.
0	**Task unattempted/partially attempted** Not enough language to make an assessment, or under 20 words.

The structure of the scale shows a number of inconsistencies. There are six bands altogether (0-5), out which there are only four bands (0, 1, 3, 5) which come with brief summary statements relating to the level of achievement. The two bands that are left empty, containing no descriptors for the level are 2 and 4. The scale does not provide the summary statements for the level of achievement either. There are descriptors for bands 0, 1, 3 and 5, with a short summary statement at the top. The descriptors of task achievement for these bands, except 0, are made up of five subcategories: (a) intention, (b) instructions, (c) effect, (d) outcome, and (e) content. They are all related to achievement, but they

observe different components of the writing process and the writing product. Subcategories (a) intention and (b) instructions are implicated in qualities outside the text. Intention refers to the disposition of the writer whereas instructions refer to the task. The other three categories (c-d) apply internally to the writing product itself. Furthermore, although band 0 looks similar to the other three qualified bands on the surface, starting with the short summary statement, its linguistic structure differs from the other three (1, 3, 5). The descriptor of the band is not divided into five subcategories: it does not provide the qualities of the writing process or the product but merely states the lack of the product.

We may observe further inconsistencies in the actual wording of the scale. Although the definition of the five subcategories implicate coherence in rating, it is not defined how the points are awarded. It is left to the raters' individual decisions which of the five subcategories they observe, which of them they consider more important when awarding a score, weakening the reliability of their scoring. In addition, the language of the descriptors is too vague to result in valid decisions. As an example, we may take the criterion of *instructions.* The band descriptors raters may observe are "all important ones followed" for 3 points and "many not followed" for 1 point. In addition to the lack of specification for "important instructions", the band for 2 points does not have any descriptors at all. The judgement of the rater about the extent to which the test takers was following instructions is definitely left to subjective impressions. This high level of subjectivity is the major characteristic of the scale and it clearly blurs the criteria of assessment of writing products and calls for improvement.

The rater interviews I carried out (Chapter 7) already unveiled that there is no consensus among raters on the interpretation of descriptors, and that the concepts of the construct need to be clarified in order to reduce the impact of specific and unwanted examiner behaviour. It became clear to me that no matter how experienced the raters are, they have diverse and contrasting perceptions of the task and the rating scale. Moreover, their understandings are often contradictory and inconsistent. This phenomenon is confirmed by the live administration results reports (Euroexam, 2018a, 2018b) which are discussed in detail below. Test scores on the writing paper usually show very little variance and within narrow limits, which does not conform to test taker performance on other test papers, from which we may conclude that the scores do not reflect test takers' writing skills but rather the way, the tacit expectations raters deploy to assess them. To support my hypothesis, I used two sessions from live test administrations for development purposes. I

chose the May 2018 administration for large sample field testing because the number of C1 test takers was the highest on that occasion (N = 584) and I opted for the July 2018 administration for development purposes based on the suitability of the essay task in that particular paper. As pointed out earlier, Task 1 of the C1 level General English test also contains a transactional writing task, whereas Task 2 contains three discursive genres the test takers may choose from. I decided on the suitability of the essay task on the basis of the Euroexam Academic topic list, I complied in Stage 1 (Appendix 4).

When looking at the July 2018 results report (Euroexam, 2018), it was evident that (a) raters prioritise certain scores while they hardly ever use others and (b) the understanding of the assessment criteria of the same writing product between the two independent raters is strongly diverse. I calculated the mean difference of the scores awarded by first raters and it was 19.93%; and between second raters and it was 17.67%. As opposed to this, the mean difference among genres was only 2.47%. It is obvious that there was a strong rater effect in the assessment of Task 1 and Task 2 of the writing paper. I also used the raw data of the same administration and checked the frequency distribution of the results. As for score distribution, Figure 12 shows that the raw scores do not show normal distribution for the essay task.

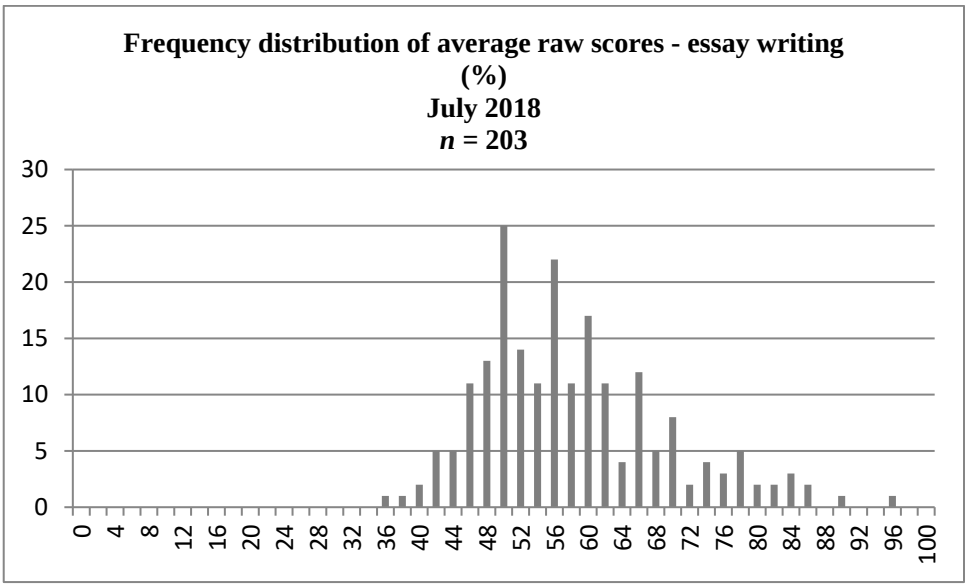


Figure 12. Frequency distribution of average raw scores – July 2018, essay task

Figure 12 clearly displays a central tendency of low score variance. We can also observe that only 1.97% of the test takers scored below 40%, and nobody scored below 36%. The peak is at 60%, which is the centrally defined pass mark for the writing paper. When we look at live administration data from the May 2018 administration, we may observe a

similar distribution for the writing tasks, making the two sets of data equally suitable for data analysis and development purposes.

A further noticeable phenomenon is the difference in score distribution of the Reading and the Writing paper. When we compare the reported scores of the Writing paper (both tasks), and the Reading paper (Figure 13 and 14), it is clearly discernible that the distribution is different.

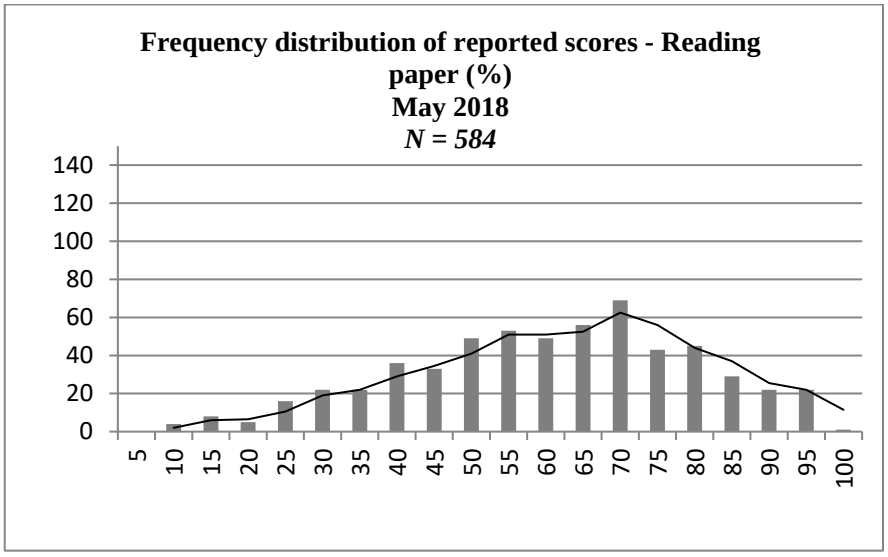


Figure 13. Frequency distribution of reported scores – May 2018, Reading paper

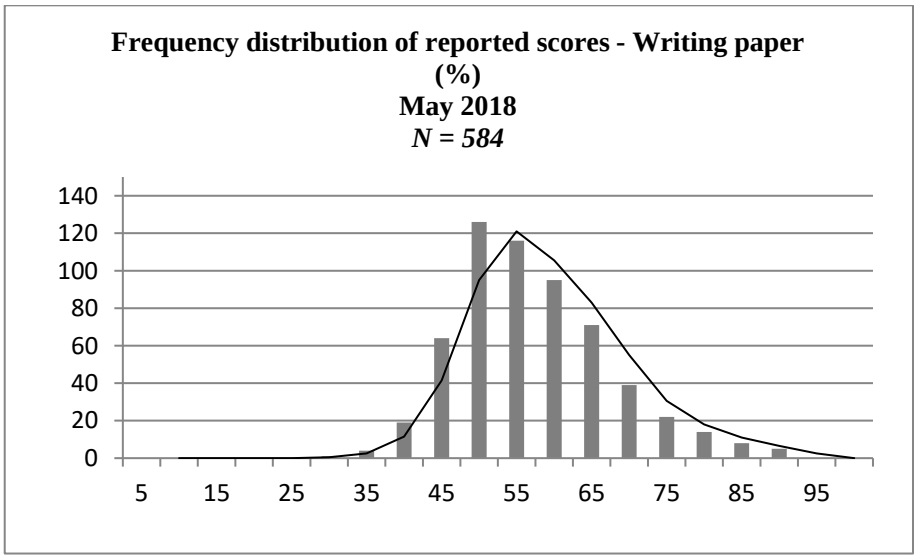


Figure 14. Frequency distribution of reported scores – May 2018, Writing paper

Since the reading tasks of the Written paper are objectively marked with an answer key, it may be concluded that that the central tendency in writing assessment is not a characteristic of test taker performance. The flat trend line in the distribution of the

Reading shows no outstanding scores, whereas the subjectively marked Writing tasks of the same test takers display low variance with no scores below 40%, and with scores grouped in the middle.

8.2 Methods

Lukácsi (2018; 2020) reports on the steps of the development of the B2 level writing tasks of the General English test of Euroexam International. After reviewing the fallacies of writing scales in general and the inadequacy of the B2 level Euroexam scale in particular, he describes the method and procedure of the development and validation of a writing checklist. The checklist items were developed based on CEFR descriptors and the detailed analysis of test taker performance. According to him, the 36 dichotomous items of the rating tool made the assessment of the writing products more reliable by distinguishing between weak and strong candidates who have previously been mistakenly classified by the central tendency of the Euroexam rating scale. The present C1 level checklist project for essays follows the major points of the research design in Lukácsi (2018; 2020) and Struthers et al. (2013) but takes into account the specific feature of the C1 level test and the genre of the discussion essay.

As outlined in Chapter 5 (Table 2-3), Stage 4 of the test development process is an independent research project with the aim of developing a task and level-specific checklist-based rating tool for the essay writing task. Stage 4 consists of two main parts, document analysis (phase 1) and empirical research (phases 2-8). The document analysis involved reviewing the relevant scales of CEFR (Council of Europe, 2001) and *The CEFR Companion Volume with New Descriptors* (CV) (Council of Europe, 2018) and the rating scales of other Academic tests of English at C1 level available online (four Academic test of international exam centres altogether). The empirical research was divided into two parts, with the first one (phases 2-6) focusing on designing and developing the items of the checklist, and the second (phases 7-8) aimed at exploring the relationship between scale-based and checklist-based scores.

At this point of my research project, the Academic test had undergone the pretesting phase (Chapter 7), but I had no live data available. The Euroexam Academic Test may be considered a profile of C1 General English Test, therefore I arrived at the conclusion that carefully selected live test results are suitable for my development purposes. The Academic pretesting had already taken place, but I did not find the size of

the sample, (N = 136) large enough to serve as a database for the development project. Fortunately, I was able to use the data sets from a live administration of the Euroexam General English Test since two thirds of the target population falls in the age-range of 14-24 (Euroexam, 2018a; 2018b), which is exactly the same target population of the Euroexam Academic Test. Consequently, I reviewed all the administrations of the Euroexam C1 general test in the 2018 season and decided to use the essay task of the July 2018 administration (N = 420), as the topic of the essay was *Getting a university education is no longer a guarantee for success* and that is suitable for the Academic validation project. Apart from the topic of the essay, the number of test takers who opted for the essay task was also satisfactory (n = 203). In addition to this, I used the May 2018 administration for the large sample field test (N = 584), where the larger number of test takers who opted for the essay task (n = 273) was more suitable for data sampling.

I divided the development project into two main phases. The first phase focused on the development of the checklist-based tool, while the second phase involved field testing with a higher number of raters and scripts. In the first phase, I worked with experienced teachers and raters (N = 4) from outside of Euroexam International. As a result of the rater think aloud protocols with Euroexam International raters (Chapter 7) namely, that they were unsure of the content and formal requirements of the discussion essay within the academic domain, I decided to consult teachers who have experience with Academic writing, university assessment and accredited language tests. Three of my four informants were university instructors, two of them are also trained, accredited raters of both international and Hungarian language tests, and one teacher was an instructor and official assessor of the International Baccalaureate Organisation. On average, they had almost 20 years (M = 17.25; SD = 6.85) of teaching experience, and 10 (M = 10; SD = 3.91) years of rater experience. Having finalised the checklist after Pilot 2, I carried out a large-sample test to compare scale-based and checklist-based scores. For rating, I used all the accredited raters of Euroexam (N = 9). Five raters were asked to use the accredited scale, whereas four were told to use the checklist-based assessment tool I developed. I chose to work with Euroexam raters in this phase deliberately to secure a control group and see how they relate to the new rating tool and to model a future live administration use. The nine raters had more than 10 years (M = 14.25; SD = 1.07) of rater experience, while their teaching experience greatly varied (M = 23; SD = 11.95).

8.3 Document analysis

The document analysis started with reviewing literature in connection with rating written products. The discussion of the fundamental works can be found in Chapter 2. In the current chapter, I focus on the documents that are directly related to the development of the checklist: the 'Can Do' statements and the illustrative scales of the CEFR (Council of Europe, 2001) and the new scales and descriptors of the CEFR Companion Volume (Council of Europe, 2018). I also review rating tools of C1 level written products in the Hungarian and international context.

The CEFR (Council of Europe, 2001) was published following 10 years of development, and is still considered to be the most frequently used framework for teachers, learners and testers inside and outside Europe. Despite the thorough and precise drafting and piloting process preceding its publication, the CEFR has been criticised for a number of reasons by different scholars (see Chapter 3). The Companion Volume (CV) (Council of Europe, 2018) can be regarded as an extension to the CEFR in response these critical points. The main focus of the CV was to update the CEFR illustrative scales and complement the 2001 edition with new scales for mediation and pluricultural competence. According to the formulation of the CV's purposes (p. 42), the volume wishes to provide descriptors relevant to a particular context, and which assessors and researchers may adapt for their purposes. In the first stage of the checklist development project, I consulted the relevant illustrative scales for written production and interaction in the CEFR and the revised and new scales of the CV. I used the CEFR and the CV parallel to be able to compare and contrast the descriptor scales for B2+ and C1 levels.

Concerning written production, the CEFR document provides illustrative scales for (a) Overall written production, (b) Creative writing, and (c) Reports and essays. The scale for Overall written production is extended in the CV. In addition to the 2001 definition of the quality of C1 level writing, the CV defines the requirements in terms of structure, conventions, tone and style of the genre. This complement seems useful for the essay checklist project, as one of the aims of the new assessment tool is to increase raters' genre awareness. The Creative writing scale is rather detailed in the CEFR, the C1 level descriptor focuses on structure, highlighting salient issues and detailing supporting points. Similarly to the Overall written production, the CV added one extra point about the structure and conventions, style and tone of written genres. Since the Euroexam essay writing task is an individual writing task, and the specifications require test takers to write

from knowledge and experience, the Creative writing scale is highly relevant for our purposes. The CEFR gives a broad definition for C1 level Creative writing skills, the main focus of the descriptor is the quality of the text, such as being imaginative, using a personal, natural style, whereas the CV specifies the genres where creative writing might be important, such as reviews and literary criticism. As for Reports and essays, the CV specifies the relevant genres (short reports and posters, to complex texts which present a case), the C1 level descriptor broadens the list of genres by adding "longer report, article or dissertation on a complex academic or professional topic" (Council of Europe, 2018, p. 77).

Originally, the CEFR failed to provide descriptors for C1 level production strategies, (a) Planning and (b) Compensating. The C1 level band for Planning in the CEFR does not have its own descriptor, but it is defined "as B2". Similarly, the scale for compensating skills directs us to consult level B2+ for the C1 descriptor, but the CEFR does not define B2+ either. The new CV provides definitions at each level for both scales. As for Planning, the CV complements the scale by the ability of adapting to different conventions, while descriptors for Compensating are now available for B2+ and C1. The difference lies in the number of gaps in the product and the ability to cover them. B2+ compensation strategies are about covering gaps, whereas C1 highlights that the purpose of the effective use of circumlocution is not only covering gaps but creative vocabulary use.

Although essay writing does not entail interaction between different parties, one of the written interaction scales is relevant for the assessment. Correspondence at C1 level is defined in the CEFR as "[the language user] can express him/herself with clarity and precision in personal correspondence, using language flexibly and effectively, including emotional, allusive and joking usage" (p. 83), which is complemented by modes of correspondence in the CV, such as personal response, and emotional allusive usage. As the specifications in the Euroexam define the task purpose as giving a personal response to academic topics, and writing from knowledge and experience, it is important that the test taker be able to give their situational underpinnings in connection with the topic defined in the task. In this respect, turntaking is also a relevant interaction strategy, it is important to see to what extent the language user is able to obtain the discourse initiative.

The components of communicative language competences are (a) Linguistic, (b) Sociolinguistic competences, and (c) Pragmatic competences. These aspects of communicative language use, as pointed out in the CV, are "always intertwined in any

language use; they are not separate ‘components’ and cannot be isolated from each other” (Council of Europe, 2018, p. 130). Despite the intertwined nature of the three components, I turned special attention to finding the most relevant ones for the assessment of a C1 level essay, in order to provide unambiguous descriptors to Euroexam raters to give common grounds for understanding, and increase their awareness of the features of the genre of the essay.

In addition to the scale for General linguistic range, there are further relevant subscales to linguistic competence, such as (a) Vocabulary range, (b) Grammatical accuracy and (c) Orthographic control. The C1 level descriptors for all three contain the following criteria: *consistent, rare errors, occasional slips, less common words.* Both the CEFR and the CV provide one scale for Sociolinguistic appropriateness. Similarly to the scales of Linguistic competences, C1 level products are described with the use of idiomatic expressions and flexible language use. The CV further elaborated the B2+ and C1 levels with appropriate language use and the ability of framing critical remarks.

The relevant scales of Pragmatic competence are (a) Flexibility, (b) Turntaking, (c) Thematic development, (d) Coherence and cohesion, and (e) Propositional precision. As for flexibility, the C1 level descriptor is new in the CV, which in addition to the level of formality, focuses on positive impact on the audience, advanced vocabulary and word order, the ability of the expression of degrees of commitment, confidence or uncertainty (Council of Europe, 2018, p. 139). Turntaking as a discourse competence is the same as the one we saw under interaction strategies. Thematic development together with Coherence and cohesion were extended in the CV with details that help define a well-built and structured writing product, such as paragraphing, the ability of using main and supporting points, and the controlled use of a variety of cohesive devices. In terms of Propositional precision, which is about the ability to “formulate what one wishes to express” (Council of Europe, 2018, p. 143), the key criteria are the degree of precision in producing language, and the ability to qualify the information given by the language user. For this reason, the CV was complemented with effective the use of linguistic modality.

In addition to modifying existing scales, the Companion Volume provides scales for competences which were defined and used by the CEFR without specifying relevant illustrative scales. The CEFR discusses plurilingual and pluricultural competence in Chapter 8, but apart from turning attention to how to build on language learners’ pluricultural repertoire, the document does not provide a detailed scale. The Companion

Volume, however, devotes a separate chapter to plurilingual and pluricultural competence, and presents a very detailed illustrative scale with descriptors from level Pre-A1 to C2.

The relevant levels for the assessment of the C1 academic writing products are B2+, C1 and C2. The different bands focus on the extent to which the test taker may be capable of reflection to socio-cultural and pragmatic differences. The key words that appear in the descriptors are: *reflection, interpretation, critical review, constructive reaction and social and cultural awareness.* These ideas significantly contribute to the development of an assessment tool in an international context, and may help assessing the test taker's awareness of cultural issues and can serve as a point of assessment when reviewing the rating tools of academic exams in the Hungarian and an international context.

As I have established, there are no accredited Academic English tests in the Hungarian system, instead, candidates may choose the products of internationally recognised test providers (IELTS, Pearson Academic, TOEFL iBT, and the Cambridge examination suites). These English language tests are not accredited in Hungary but may undergo a nostrification process (referred to as "nationalisation" by the Educational Authority (Educational Authority, 2020c). I consulted the websites of the above listed academic tests and looked at the assessment tools and their criteria to see how they assess the writing products, what criteria and band descriptors they use.

The IELTS Academic writing scale (IELTS, 2020) uses a nine-band scale with four criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy. The scale does not contain empty bands, there are descriptors in each cell of the assessment matrix. Task achievement concentrates on the requirements of the task, which are listed in bullet points within one band. Apart from task requirements, the criterion also includes purpose and appropriate tone for most bands, except Band 8 and 9. The scale unites the assessment of coherence and cohesion, and this criterion focuses on logically organised information and the use of cohesive devices. The descriptors define four body paragraphs, and the requirement of correct referencing throughout the text. The criterion of Lexical Resource uses quite vague descriptors, such as *adequate, sufficient, skilful* use. Although an IELTS score may have a significant impact on the test takers' studies, it is not clear how a rater differentiates between *adequate* and *sufficient* use of lexical resources, which appear in two different bands. The scale is more precise in Band 8 and 9 which provide examples for *skilful* use – uncommon lexical items, collocations. As

for Grammatical range and accuracy, the assessment scale puts great emphasis on the use of sentence structure, at the same time the definition of errors for the different bands is somewhat vague (*frequent errors*, *some errors*, *many error free sentences*.

Pearson Academic (PTE) uses integrated tasks (reading into writing, listening into writing) and independent tasks to assess writing (Pearson Academic, 2019). Since the Euroexam Academic only uses independent tasks, the assessment criteria of the integrated tasks of PTE are not reviewed here. The independent essay writing task of PTE is both machine-scored and human rated, the test centre uses "an automated scoring tool that is powered by Pearson's state-of-the-art Knowledge Analysis Technologies™ (KAT™) engine." The Scoring Guide provides the rating scales for the traits assessed. These are, together with the raw scores are presented in Table 15.

Table 15
The Scoring of the Seven Traits of Essay Writing. (Pearson Academic, 2019, p. 39.)

Traits	**Maximum raw score**
Content	3
Form	2
Development, structure and coherence	2
Grammar	2
General linguistic range	2
Vocabulary range	2
Spelling	2
Maximum item score	**15**

The assessment tool uses three or four bands depending on the maximum raw score of a criterion. Form and spelling are only machine scored as these traits can be counted and objectively scored. Form defines the length of the text and spelling defines the number of spelling errors for each band. Content is rated based on the number of points the essay covers, which is both human rated and machine scored, with the numbers for the different bands are clearly defined. Development, structure and coherence focuses on logical structure and the number and quality of linguistic devices the text displays to link paragraphs. The criterion of Grammar is the only one that contains some vague language, e.g. Band 1 is defined as *relatively high degree of grammatical control* as opposed to *consistent grammatical control* in Band 2. Consistent, however, does not mean error free, because rare errors are accepted in Band 2 as well. The three bands of General linguistic range are characterised by the use of the descriptors *basic*, *sufficient* and *mastery*, which might be difficult to decipher for human raters; whereas Vocabulary is more precise, it

focuses on the domain of academic topics, and the command of idiomatic expressions. Although subjectivity is present in the assessment tool, it is overridden by the fact that two traits are only objectively assessed by machines, and one trait is assessed by the help of artificial intelligence and human raters.

TOEFL iBT assesses writing skills with integrated and independent tasks (Educational Testing Services, 2019). As for the independent task, there are six scoring bands from 0 to 5, and they represent different levels. 4 statements for bands 3 to 4: (a) topic and task, (b) organisation, (c) progression and coherence, (d) language and vocabulary. Band 2 has a fifth descriptor about the number of errors, whereas Band 1 contains 3 descriptors (organisation), (b) task fulfilment and (c) number of errors. Band 0 has only 1 descriptor, but unlike the Euroexam scale, it describes the product as "merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank", which is a more objective criterion.

The Cambridge writing assessment scales (University of Cambridge Language Examination Services, 2016) are divided into six bands from 0–5. They share a feature with the Euroexam scales as they lack descriptors for bands 2 and 4, also Band 0 is defined as "below Band 1". Descriptors are provided for Bands 1, 3 and 5, while the "Band 2 descriptor indicates a performance which contains features of Bands 1 and 3, and Band 4 indicates features of Bands 3 and 5" (p. 2). The four criteria used are (a) Content, (b) Communicative achievement (c) Organisation and (d) Language. Content focuses on relevance and informing the target reader, which also appear in communicative achievement. As for Organisation, the three band descriptors are almost identical, which might make assessment challenging. The descriptors for Language are the most detailed of the four criteria, they contain descriptive elements for grammar, lexis and the number of errors. Similarly to the IELTS scales, the main descriptor of high quality writing products is the use of uncommon lexical items and being error free.

I also consulted research papers that present the full development process of checklist-based rating tools for the assessment of writing. Harsch & Seyferth (2020) took the initiative to design a checklist for a local university test. Since they see the advantage of the checklist in their suitability for judging the specific goals and targets, they designed a tool in which they incorporated the assessment of achievement and proficiency tests (Figure 15). The collaborative development project used the CEFR for their basis and

relied on instructor experience as the checklist was designed for 21 languages. The paper uses the criteria of Task fulfilment and Organisation for illustration. The checklist draft contains well defined statements concerning the particular dimensions of the task, however, we can see that the assessors may choose from five categories: (a) Completely fulfilled, (b) Almost fulfilled, (c) Largely fulfilled, (d) Partly fulfilled and (e) Not fulfilled. It is clearly discernible that the checklist does not contain dichotomous statements, but it may only be used as a horizontal scale without the advantages of a checklist with yes/no items.

Task dimension		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Task fulfilment	Can complete the **task** in terms of content and writes 80-100 words **about:** *[specify topic]*					
	Genre *[specify requirements]*					
Linguistic dimension		Very good ++	Good +	Satisfac-tory +/-	Sufficient -	Insufficient --
Organisation	Can write a series of simple phrases and sentences and link them with simple connectors such as *and, but, or, because, when, if*.					

Figure 15. Initial checklist draft, level A2.1. (Harsch & Seyferth, 2020, p. 7).

Due to shortcomings of the draft checklist, which are largely similar to scale-based assessment, the researchers had to drop the first item on text length after trialling: it was impossible for them to decide what the different categories meant in terms of word count. In addition to this, in the course of calculating the results, the five categories were encoded as numbers from 1 to 5, which made the use of the checklist identical to that of a scale. For these reasons, this particular checklist development project was irrelevant for my purposes.

The research of Struthers et al. (2013) presents the development process of a checklist to assess cohesion in children's writing. They followed the steps of Crocker and Algina (2006) for the development of their assessment tools. The pre-testing of the checklist showed that the assessment tool can be used with high inter-rater reliability and it is able to discriminate between weak and strong writers. They also provide their preliminary cohesion checklist in an appendix to the paper (Figure 16) with a list of well-formed dichotomous statements.

Cohesive marker	YES	NO
1. All pronouns refer to some previously mentioned noun.	1	0
2. All pronouns have a referent in the previous sentence or clause.	1	0
3. All demonstratives (e.g. this, these, that, those, here, there) have a clear referent in the previous text.	1	0
4. Referents for nouns used with 'the' have an unambiguous previous mention in the text or can be inferred from world knowledge.	1	0
5. Each sentence is connected to the one preceding it by at least one form of reference.	1	0

REFERENCE SUB-SCORE _____

Figure 16. Preliminary Cohesion Checklist (Struthers et al., 2013, Appendix B)

The raters in this model are not left any room for interpretation, as the descriptors are explicitly qualified: either *all pronouns, all demonstratives, each sentence* and *at least one form of reference* or none. This will lead to high inter-rater reliability and will increase the validity of the assessment tool as it leaves no room for test taker human fallacy.

8.4 Teacher task completion: immediate recall

After reviewing the relevant literature, the next stage of the development project was item development based on empirical data. As pointed out above, I used the live data from the July 2018 administration. In this administration, the number of test takers who chose the essay for the optional writing task accounted for approximately half of the population. The topic of the essay was: *Getting a university education is no longer a guarantee for success*. The test takers did not have an input text to rely on, they had to write from knowledge and experience. The task instructions only specified requirements in relation to structure and coherence: *Explain your points for and against; and arrive at a conclusion at the end. Make sure you state your argument in a logical way.* The four teachers were requested to complete the task themselves to see what they themselves think of a good quality essay in practice. The instructions they received were to follow the task rubric and to model the time allowance of test takers, they were asked to write on paper and not to spend more than 40 minutes on writing. After task completion, I used the immediate recall technique of verbal protocols with them to identify how they approached the writing task and what they paid attention to that resulted in a good quality essay. The verbal protocols were recorded, transcribed and analysed with MaxQDA – the same procedure I followed in the interview data in Chapter 6 and Chapter 7. In the 985-word transcript, I could identify four thematic categories: (a) time spent on task, (b) what needs to be considered before and during writing, (c) what makes a high quality essay, and (d) linguistic features of a high quality essay. I identified topic (a) and (b) as elements of the writing process and (c) and (d) that of the quality of writing. The results of the thematic analysis are displayed in Table 16.

Table 16
Thematic Categories in Verbal Protocols of Teacher Task Completion

	Thematic category	**Emerging elements in answers**
writing process	time spent on task	30 to 60 minutes, too little time, more time needed
	what needs to be considered before and during writing	planning, readability, importance of title, paragraphing, paragraphs of equal length, fit on one page,
quality of writing product	what makes a high-quality essay	balanced arguments, formal but not impersonal style, topic sentence in each paragraph, supporting evidence, example(s), more than one body paragraph, more than one sentence in a paragraph, emerging key words, creativity
	linguistic features of a high-quality essay	links between paragraphs, complex linkers, parallel structures, rhetorical forms, compound sentences

As for the writing process, we can observe in Table 16 that the allocated time (approximately 40 minutes) proved to be too short even for experienced teachers. One of the informants admitted having difficulties with completing the task in 40 minutes:

> **Text: Teacher_3**
> **Code: time spent on task**
>
> It was very difficult and very time consuming to write the essay. The only reason for this is that one does not write essays as a teacher.

Another teacher was wondering whether students who are preparing for an exam count as more experienced essay writers:

> **Text: Teacher_4**
> **Code: time spent on task**
>
> Since I know there are 60 minutes for the writing assignment [two tasks], I had to make sure I could fit in like 40 minutes. I think you need to be extremely skilled at writing for this. I know this is an exam and the skill of being able to write an essay in a given time is part of it, but I think the writing time should be increased.

Since increasing the time allowance in case of an existing test portfolio is a managerial decision rather than a professional one, I could not consider this result any further in the course of the development project.

All four teachers seem to have agreed in the qualities of a well-formed essay, their comments in the verbal protocols were coherent and supported one another:

Text: Teacher_1
Code: what makes a high-quality essay

As I wrote, I paid attention to the length of the paragraphs - I didn't count the words, I just made sure they were approximately of the same length. I put topic sentences and logical connectors at the top of the paragraphs, tried to avoid repetition. I was looking for synonyms and was also careful to use modality. I paid attention to the structure: one intro paragraph, two paragraphs "for", two "against" and one conclusion.

Text: Teacher_2
Code: what makes a high-quality essay

I paid attention to a number things while writing… like… to make normal arguments and answer the question, to have at least two arguments on each side, to make the arguments logical by using a topic sentence at the beginning of each paragraph, and to be as consistent as possible.

Text: Teacher_3
Code: what makes a high-quality essay

What I promise in the introduction, I give in the text (balanced view). I have valid arguments for and against the statement. The conclusion summarizes the thoughts described so far, does not open a new circle of thought. The style is formal but not impersonal… like…Giving credit to sceptics, I can give…). I tried to write an essay for educated and interested readers

Text: Teacher_4
Code: what makes a high-quality essay

I always try to pay attention to the structure and the proportions: there should be a separate introductory/concluding paragraph and the body between them should be divided. The structure of the paragraphs is also important, I can only imagine one-sentence paragraphs with a very definite purpose, it is much more important to get the line of thought from one point to another within one paragraph, which can then be taken up by the next.

This experience was largely different from the one I had in the rater interviews, where the responses of the three raters deviated from each other to a great extent. I could conclude that the teacher task completion and the think aloud protocols with the technique of immediate recall fulfilled my original aim of creating a basis for a common understanding of a good quality essay.

In order to establish whether the teacher comments could serve as a basis for the statements of the assessing tool, it was essential to assess test taker performance using the same criteria. For this purpose, I chose three test takers scripts (Appendices 10-12) with targeted sampling (Wagner, 2010, p. 30) to represent three different performance levels: fail, pass and pass with distinction. I used the raw data of the July 2018 administration and calculated percentages from the raw scores awarded for this one task. After that I divided the results into three groups, following the guidelines of calculating Ebel's range of discrimination indices (Ebel & Frisbie, 1991), and chose a script as a representative of that range. The four teachers assessed the script based on the criteria I complied based on the verbal protocols. They verbalised their rating process through think aloud protocols, which were recorded and transcribed. Based on the 2515-word transcript, I listed the characteristics of essays the teachers were rewarding and penalising during their evaluation and grouped them according to the relevant scales of the CEFR.

(1) Positive

a) orthographic control
   i. legibility
   ii. paragraphing – more than one body paragraph
   iii. title
b) thematic development
   i. statement of position, topic/thesis sentence
   ii. refutation, building contrast
   iii. summary
c) creative writing
   i. thick description of context
   ii. genuine ideas, creativity
   iii. stating an opinion
   iv. self-disclosure
   v. multiple points of view
   vi. balanced argument
   vii. presenting unexpected content
   viii. use of rhetorical questions
   ix. offering examples/recommendations
d) coherence and cohesion
   i. cohesive devices (secondly, despite, on the other hand)
   ii. links/transition between paragraphs

e) grammatical range
    i. complex structure
    ii. complex, compound sentences
f) grammatical accuracy
    i. varied use of modal verbs
    ii. correct word order
    iii. correct referencing
g) vocabulary control
    i. collocations
    ii. strong lexical items
    iii. idioms
    iv. proverbs

(2) Negative

a) orthographic control
    i. illegibility (endnotes)
    ii. spelling (apostrophe, different meaning, extensive use of exclamation marks)
b) thematic development
    i. not enough detail
    ii. deviate from subject
    iii. counterargument/refutation missing
    iv. listing (instead of arguing)
    v. poor contextual coverage
    vi. vague/unclear argumentation
c) sociolinguistic appropriateness
    i. genre different than expected
    ii. inappropriate style/formality
d) coherence and cohesion
    i. illogical development (topics merge)
    ii. poor coherence, reader has to re-read to follow (logic)
    iii. reference mistake
    iv. cohesive tie missing
e) grammatical range
    i. simple sentences
    ii. lack of structural range (grammatically simple)
f) grammatical accuracy
    i. varied use of modal verbs
    ii. correct word order
    iii. correct referencing
g) vocabulary control
    i. lexical mistake (wrong word)
    ii. lexical mistake (non-existent word)
    iii. lexical mistake (part of speech)
    iv. lexical mistake (idioms used out of context)
    v. repetition (lack of synonym)

On the basis of this organisation of the characteristics that emerged in the verbal protocols, we may conclude that the findings of the review of relevant assessment literature (CEFR,

CV, assessment tools of other Academic test providers) and the empirical data reinforced each other. Consequently, the list of positive and negative characteristics could serve as a basis for the design of a preliminary checklist.

The next step in the development process was to reformulate the characteristics listed above and the descriptors of the relevant CEFR scales and design dichotomous statements for a checklist-based rating tool. Based on the above list of qualities, the Euroexam writing construct, and fifteen relevant CEFR scales of communicative language activities and strategies, I designed a preliminary checklist. The scales in alphabetical order are (a) Coherence and cohesion, (b) Creative writing, (c) Correspondence, (d) Flexibility, (e) General linguistic range, (f) Grammatical accuracy, (g) Orthographic control, (h) Overall written production, (i) Planning, (j) Pluricultural repertoire, (k) Propositional precision, (l) sociolinguistic appropriateness, (m) Thematic development, (n) Turntaking, and (o) Vocabulary control.

8.5 Preliminary checklist use

The preliminary version of the checklist contained 34 items in the form of statements as statements are said to be easier for the raters to use than questions (Struthers et al., 2013). In addition to the statements, the checklist contained the name of the relevant CEFR scale and a number of concept check questions to make sure raters have the same concepts in mind when making a binary choice decision. An excerpt of the checklist the raters used for their first rating is displayed in Figure 17 the full checklist is provided in Appendix 13.

The logic of the statements is not punishment or reward but merely noticing the presence or absence of a feature.
The reference scale headings in the middle serve as a direct link to the CEFR (2018).
The concept check questions on the right serve to ease the decision-making process.
If the answer to all the questions in a cell is in the positive, allocate a 1 indicating that the target trait is present. If there is a negative, answer, allocate a 0.

Statement	CEFR reference scale	Concept check questions
1. This text is **legible**, i.e. the reader doesn't have to guess what the writer is trying to say.	orthographic control	Can you read the text without having to re-read words? Is the text legible? Can you keep your role as reader?
2. This text follows the standard **layout** of an essay.	orthographic control	Does the script follow standard paragraphing conventions? Are there at least four visible paragraphs (intro, more than one body paragraph, conclusion)?
3. This text is **clear** and concise.	overall written production	Are there signs of planning so that the reader's work is easier? Can you keep your role as reader? Can the writer employ the structure and conventions of the genre?
4. **Spelling** is consistently accurate.	orthographic control	Are there only **two or fewer** spelling mistakes?

Figure 17. C1 level checklist for essays – Preliminary version

As the instructions at the top of Figure 17 make it clear, the main idea of the checklist is that none of the items on their own fail or pass a test taker, the statements focus on the

presence or the lack of a characteristic feature the teachers identified earlier, and which are also linked directly to the CEFR. It is also made clear that awarding a point is possible in case all the answers to the concept check questions are in the positive. This way, the traditionally subjectively assessed writing product becomes as non-biased and reliable as possible.

The teachers assessed the first three scripts of the targeted sample, representing three typical qualities of writing, fail, pass and pass with distinction. The purpose of the preliminary trial was to reveal, along the lines of Research Question 3, how the teachers use the checklist and what they think about the new rating tool. In addition to this, in order to test the objective nature of the statements in the new checklist design, the four teachers were also asked to support their decisions with evidence from the texts. After their first checklist-based rating experience, I invited the four teachers to carry out verbal protocols to find out (a) how they felt about the rating tool, and (b) what they found easy or difficult, and (c) any suggestions they may have.

The teachers felt generally positive about the rating tool, they found it easy to answer the questions and decide if a statement falls in the positive or in the negative category. As for their recommendations, they suggested a change to the order of the questions and less descriptors. The typically emerging elements of the verbal protocols are displayed in Table 17.

Table 17
Thematic Categories Based on Teacher Verbal Protocols after Preliminary Checklist Use

Thematic category	**Examples from teacher answers**
Feelings about rating tool	"I liked rating with a checklist, the concept check questions are really useful"
	"I liked that it is more about discourse and composition features rather than grammar"
	"I felt more positive and confident about my rating"
	It was a good experience, although I was sceptical at the beginning of the project"
	"My rating became more balanced"

Thematic category	Examples from teacher answers
	"It's a great tool to give feedback to the test taker on their writing quality"
	"The criteria are objective; I knew what I was looking for"
	"I think the checklist may help raters who rate a large number of scripts because it does not let you award scores based on gut feelings"
Easy	"It was easy to differentiate between good and bad essays"
	"The checklist helps focus my attention"
Difficult	"Difficult to scroll up and down"
	"Difficult to answer all questions"
Suggestions	"Maybe we could change the order of the items so that spelling could come after punctuation"
	"There are too many items"
	"Too many concept check questions"

It was clear from the interviews that the teachers were more positive about the new rating tool than negative and did not report about extreme difficulties. As for the evidence they provided, it showed that the concept check questions could successfully direct the teachers towards elements of language and discourse that may determine the quality of a writing product.

Based on their suggestions concerning the number and order of items, I amended the checklist to make it more user friendly. (See the amended checklist in Appendix 14). I changed the order of the items so that the items that belong to the same category come after each other. As all the four teachers found Items 8-10 in the preliminary list repetitive, I deleted Item 10. Item 12 was partially cut; its concept check questions were moved to Item 19. The teachers also found item 23 too vague, so it was split to two separate items targeting tenses and aspects and the use of pronoun forms. This way the second version of the checklist had 33 items.

8.6 Pilot 1: rater agreement and reliability

After redesigning the checklist, I designed a small sample pilot test using the same population of the July 2018 administration. The method of sampling was non-proportionate stratified random sampling (Mackey & Gass, 2005, pp. 119-124). The population was divided into 3 tertiles, and altogether 16 scripts were chosen, 8 from the lower and 8 from the upper ones. The sampling was non-proportionate because the number of scripts in one tertile did not reflect the proportion of the population. The reason for this was the low number of low performer scripts (as we saw above in Section 8.1, there were no test taker results below 34%). This way the upper and lower tertiles represented the high performers and the low performers of the population.

The teachers were allocated 16 common scripts and after the rating procedure, a rater conference was held in the course of which they could discuss the scripts and could modify their original ideas. This modification, however, did not affect their original rating. They went through the items of the checklist together and came up with an agreed mark for each statement. This way a consensual score was formed that I call and introduce as Rater 5 in Table 18. In the first pilot phase of the development project, due to the low number of scripts in the sample, item level statistics were not computed. The main aim of this phase was to deepen the understanding of the checklist items and enhance rater agreement. I examined rater agreement and reliability with methods of Classical Test Theory (CTT) and used IBM SPSS 24 for analysis. The correlation between rater scores and the agreed score is displayed in Table 18.

Table 18
Correlation between Raters – 16 common scripts

		Rater_1	**Rater_2**	**Rater_3**	**Rater_4**	**Rater_5**
Rater_1	Pearson Correlation	1	.829**	.891**	.878**	.979**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	16	16	16	16	16
Rater_2	Pearson Correlation	.829**	1	.780**	.710**	.878**
	Sig. (2-tailed)	.000		.000	.002	.000
	N	16	16	16	16	16
Rater_3	Pearson Correlation	.891**	.780**	1	.786**	.915**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	16	16	16	16	16
Rater_4	Pearson Correlation	.878**	.710**	.786**	1	.906**
	Sig. (2-tailed)	.000	.002	.000		.000
	N	16	16	16	16	16

		Rater_1	Rater_2	Rater_3	Rater_4	Rater_5
Rater_5	Pearson Correlation	.979**	.878**	.915**	.906**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	16	16	16	16	16
** Correlation is significant at the .01 level (2-tailed).						

In Table 18 we can see that there is significant high positive correlation between raters, ranging from .710 to .979. The strength of the checklist lies in using easily understandable items which describe a feature that is directly observable in the writing products. It can be proven by the correlation obtained after rater consultation, between the four raters and the agreed score (Rater 5), which is between .878 and .979. In addition to the Pearson correlation matrix (Table 18), I computed Intraclass correlation (ICC) to indicate the reliability of rater agreement. A high degree of reliability was found between the four raters. The average measure ICC was .945 with a 95% confidence interval from .882 to .979 ($F(15,45) = 18.172$, $p < .001$). Since an average measure above .75 is considered high, we may conclude that even after the first use of the new rating tool, the rating became more objective and reliable.

Along the lines of Research Question 3, I made the concept clarification questions more specific. The concept clarification questions relied on the emerging topics and questions I recorded in the course of the rater conference. A major concern of the teachers after the first pilot was the low number of items for grammatical range and accuracy. Compared to the B2 transactional writing checklist (Lukácsi, 2017; 2018), where the number of grammar and vocabulary items were 15, the version of the C1 essay checklist we used in the first pilot contained only 9 items in this category. The reason for this is twofold. Firstly, it may be explained by the nature of C1 level descriptors in the CEFR, where grammatical and lexical range and accuracy is often described as "complex" and characterised by a "wide range" and the "lack of mistakes". Further to this, C1 level entails proficient use and the ability of performing complex tasks for different purposes, which unlike at B2 level, does not allow for highlighting specific examples for correct use. The need of the teachers to enhance the objective nature of their rating was considered a valid point so after the rater conference the checklist went under an additional review process at the end of which a new item was added in connection with grammatical agreement: *This text shows full consistency in the use of grammatical agreement*. After the first pilot, this

version of the checklist (Appendix 15) was used in the following stage of the development project concerned with the reliability of the instrument (Pilot 2).

8.7 Pilot 2: reliability of the instrument

In the second pilot study, the teachers were allocated 48 scripts from the lower and upper tertiles. Each of them received 12 writing products which they assessed with the help of the 34-item checklist. In this stage the teachers did not have common scripts, each product was assigned to one rater only. Due to this arrangement, examining inter rater agreement could not have been part of the second pilot stage, but the number of scripts allowed me to compute item level statistics and investigate the reliability of the instrument itself.

In order to examine the reliability of the instrument, I compared the original reported scores, which were based on assessment with the original rating scales, with the scores based on rating with the checklist. Similarly to the May 2018 results (Figure 14 above), the July scale-based reported scores show low variance and a strong central tendency. As pointed out above, the lowest score of 38% allowed sampling based on 3 tertiles, the lower and upper ranges of which (n = 48) showed a score distribution displayed in Figure 18.

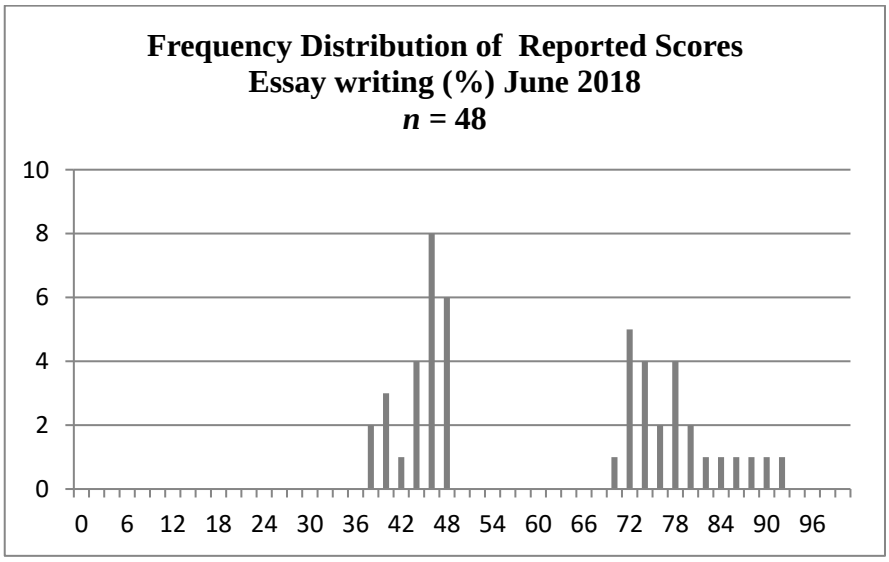


Figure 18. Frequency distribution of reported scores – low and high performers (%)

In the course of the B2 transactional email development project, Lukácsi (2017; 2018) found that checklist scores show greater variance instead of a strong central tendency. Based on this previous finding, I expected a fairly even distribution of scores both among the low performers and the high performers. The frequency distribution of the checklist-based scores for the 48-script sample can be seen in Figure 19.

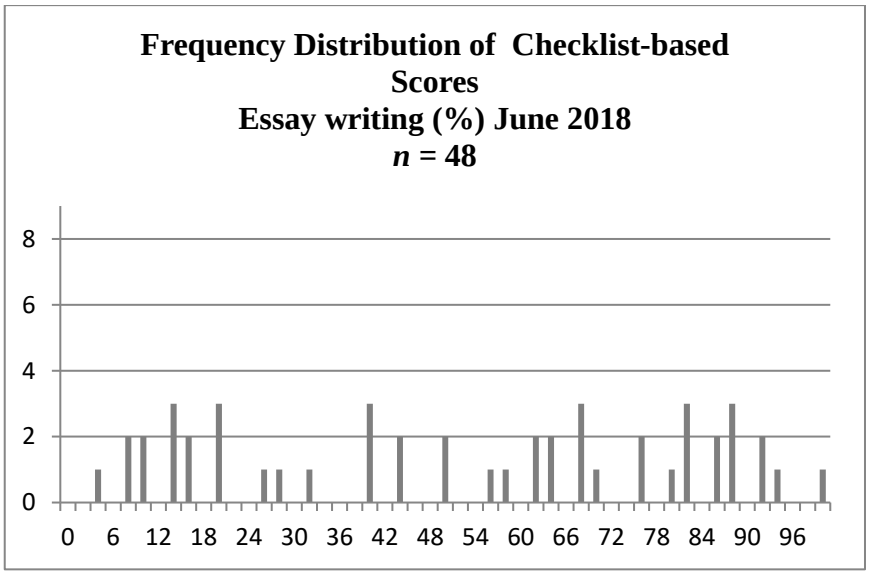


Figure 19. Frequency distribution of checklist-based scores – low and high performers (%)

Although it looks as if the scores are evenly distributed, I wanted to see whether there is a significant difference between the scale-based and the checklist-based scores of low and high performers. I carried out a paired samples t-test to compare the reported scores to the checklist scores. As for the low performers, there was a significant difference in the scores for the scale-based and the checklist-based scores (scale-based scores M = 43.88; SD = 15.84) checklist-based scores (M = 24.62, SD = 2.92,) conditions $t(23)$ = -6.12, p = .00) while the high performers showed no significant difference (scale-based scores M = 77.54; SD = 6.17; checklist-based scores M = 77.53; SD = 12.15, conditions $t(23)$ = -.07, p = .99.

In addition to comparing low and high performers, I computed item-level statistics. I used the categories of CTT, and looked at the internal consistency of the rating tool based on item total correlation (Cronbach's alpha), standard deviation (*SD*), item difficulty (*p-value*) and item quality (*Ebel's D*) of individual items. The p-value or facility value equals the mean score on an item, indicating item difficulty, where low scores imply high difficulty, whereas high scores imply low difficulty. I calculated *Ebel's D* or discrimination index to express item quality, in other words how well the item separates low performers and high performers. The reliability measure based on how closely the items of the rating tool are related to each other is α = .941, which indicates high reliability. The item level statistics are displayed in Table 19. I highlighted the values which do not fit the Hungarian accreditation requirements (Cronbach's $\alpha \geq .75$; $.70 \geq$ p-value $\geq .30$; Ebel's D $\geq .30$).

Table 19
Item-level Statistics of the 34-item Checklist

	SD	*p-value*	*Ebel's D*
Item_01	.42	.77	.13
Item_02	.45	.71	.64
Item_03	.49	.58	1.00
Item_04	.50	.56	.64
Item_05	.49	.42	.70
Item_07	.45	.29	.54
Item_08	.48	.65	.86
Item_09	.48	.65	.86
Item_10	.50	.52	1.00
Item_11	.47	.67	.86
Item_12	.50	.54	.93
Item_13	.47	.33	.62
Item_14	.50	.48	1.00
Item_15	.50	.52	1.00
Item_16	.44	.73	.64
Item_17	.44	.27	.32
Item_18	.45	.29	.54
Item_19	.49	.40	.85
Item_20	.48	.63	.64
Item_21	.50	.52	1.00
Item_22	.50	.54	.63
Item_23	.49	.40	.85
Item_24	.45	.71	.64
Item_25	.49	.60	.71
Item_26	.50	.48	.77
Item_27	.47	.33	.85
Item_28	.50	.48	.92
Item_29	.50	.44	.85
Item_30	.50	.52	.48
Item_31	.48	.63	.79
Item_32	.47	.33	.77
Item_33	.47	.33	.77
Item_34	.50	.54	.77

The figures of the item-level statistics indicate high item quality. The discrimination index is the highest possible for 5 items, and only 1 item falls below .30, it is Item 1. For this item (*This text is legible, i.e. the reader doesn't have to guess what the writer is trying to say*), we can see that the figures fall outside the range specified by the requirements. The

reason for this is twofold: legibility, which is a crucial quality of a hand-written text, is easy to achieve, however on its own it does not imply a high-quality writing product. Although the item does not discriminate well between high performers and low performers, I decided to keep it as a basic marker for readability. Apart from Item 1, there are two other items which seem to have been easy for the test takers to achieve. The p-values of Item 16 and 24 (*The writer adopts the level of formality adequate to the topic, agents, situation, domain, etc.; This text shows full consistency in the use of proforms*), are slightly above the acceptable, which suggest that these criteria are easy for most candidates, however, that discrimination indices are high for both items, which means that those who performed well on these, performed well on other items too. On the contrary, Items 17 and 18 with low p-values were challenging for the test takers. These two items (*Each paragraph presents one distinct and unified idea; Each paragraph contains a topic sentence*) concerning structure and coherence were almost equally difficult to achieve, but their discrimination index shows that low performance in this respect indicates low performance on the whole task.

Although the rating tool proved to be highly reliable with acceptable item-level statistics, a new revision round took place before the large-sample test. Based on teacher feedback, four items were deleted from the checklist, leaving us with 30 items (Appendix 16). The reasons for deleting items were both practical and professional. Most importantly, the teachers reported a 5 to 15 minute-long rating time, which they wanted to reduce; therefore, I examined the item level statistics of the second pilot and found four items which could have been either deleted or merged with other items.

Items 15 (*This text would make the appropriate effect on the intended audience*) and 10 (*The content elements required by the task instructions are elaborated in appropriate detail*) were merged based on the consideration of their identical item-level statistics, and 31 (*The style and tone of the text is appropriate*) and 16 (*The writer adopts the level of formality adequate to the topic, agents, situation, domain, etc*.) were also joined as the content of the statements were too similar. There was a strong positive correlation between items 25-26 which described complex grammatical structures (*This text shows full consistency in the use of grammatical agreement; This text demonstrates that the writer can use complex sentence structures*) and item 21 (*The text is characterised by complex grammatical structure*s) but item 21 had better item level statistics. Based on the indices of *Cronbach's Alpha if Item Deleted*, the deletions did not affect the high level

of the original reliability measure. At the end of Pilot 2, the final of the checklist contained 30 items (Appendix 16).

Based on the item levels statistics, we may conclude that the population was well prepared for the writing test in terms of formal requirements (legibility, grammar and cohesion). However, coherence and structure proved to be difficult for them. The reliability of the rating tool and the item statistics may be considered high for the population. The teachers who took part in the two-part pilot of the checklist development project understood the rating criteria and showed consistency and reliability in their rating. Their comments during the process contributed to the formulation and reformulation of the items to a great extent.

8.8 Field testing: comparing scale-based and checklist-based scores

The third round of data collection involved the highest number of scripts and raters. I used the raw data from the May 2018 live administration when 548 test takers took the writing test, out of which 273 chose the essay task. The rating of the live administration involved 5 accredited Euroexam raters, while 4 raters assessed the 120-script sample with the checklist.

The relevant phase of the B2 checklist development project that served as a model for my research also used a sample of 120 scripts. However, the number of test-takers at B2 level allowed the researchers to divide the population into quartiles and use quota sampling. In the case of C1, quota sampling was not possible for two reasons. On the one hand, the number of test takers is relatively low for dividing them into four quartiles; on the other hand, the strong central tendency and the proportion of low performers compared to the whole population undermined the design proposed by the other research project. Consequently, I decided to use simple random sampling, which gave me a sample that accurately represented the population and allowed me to draw valid conclusions about the entire population.

The data collection plan used the original rater pairs of double rating for the live administration and allocated 30 scripts each to the checklist raters using the same design (Table 20).

Table 20
The Allocation of Scripts and Raters

Scale-based rating			Checklist-based rating		
Rater 1	**Rater 2**	**Number of scripts**	**Rater 1**	**Rater 2**	**Number of scripts**
Rater P	Rater Q	65	Rater A	Rater B	30
Rater Q	Rater S	48	Rater B	Rater C	30
Rater R	Rater T	54	Rater C	Rater D	30
Rater S	Rater R	56	Rater D	Rater A	30
Rater T	Rater P	50			

This set up allowed me to examine rater behaviour for first and second raters. Inter-rater reliability and correlation for the scale-based rating was computed for all the essay scripts (*n = 273*), while I used the random sample (*n* = 120) for examining rater behaviour for checklist users.

In addition to using the checklist, the four raters who took part in the project were asked to provide an overall judgement based on holistic impression about the products on a scale of 3, where 1 meant fail, 2 indicated pass, and 3 pass with distinction/outstanding. The use of this method was motivated by two considerations. In the first place, it helped separating the low and high performers, and further differentiated between the successful test takers. Furthermore, it was based on a practical consideration. In order to reduce rating time of double rating, I wanted to see whether the two raters could use an analytic and a holistic rating tool with an equally reliable outcome. In order to do this, the reported scores of the live administration were converted to holistic scores (Table 21).

Table 21
Conversion of Raw Scores to Overall Scores

Raw scores (%)	Overall scores
0-59	1
60-84	2
85-100	3

The determination of score bands reflected the accreditation requirements, hence the pass score of 60%. This conversion allowed me to compare the overall judgement of the four checklist raters to the results of the live administration.

As the first step of the analysis of raw data from the live administration, I calculated the mean scores of the rater pairs and first counted the proportion of exact matches between the scores, then computed statistics for rater agreement (correlation) and

inter-rater reliability (Krippendorff's alpha). The statistics for rater pairs are displayed in Table 22.

Table 22
Statistics for Rater Pairs – Scale-based Assessment

Rater 1	Rater 2	*M* (R1)	*M* (R2)	Exact match (%)	*r**	α_K
Rater P	Rater Q	17.46	14.37	4.55	0.36	-.11
Rater Q	Rater S	13.65	16.06	8.16	0.56	-.64
Rater R	Rater T	18.11	20.07	12.73	0.80	.73
Rater S	Rater R	16.43	17.07	21.05	0.64	.58
Rater T	Rater P	18.48	17.32	11.76	0.71	.70

*Correlation is significant at the .01 level (2-tailed).

The data in Table 22 clearly highlights the issues with raters and scale-based rating, which were raised when discussing the rater interviews and the pretest statistics, respectively, in Chapter 7. The correlation between rater scores is lower than expected, furthermore the α_K reliability measure displays negative figures. Correlation and reliability is the lowest for the P-Q and the Q-S rater pairs, based on which we can conclude that Rater Q does not fit the rating pattern of other raters. However, as this rater had the lowest mean scores, I decided not to exclude their scores from the examination since getting rid of the only rater who shows a tendency for severity would further skew the score distribution (Figure 20). The mean scores and standard deviations for the scale-based reported scores show low variance (*M* = 55.10; *SD* = 12.64).

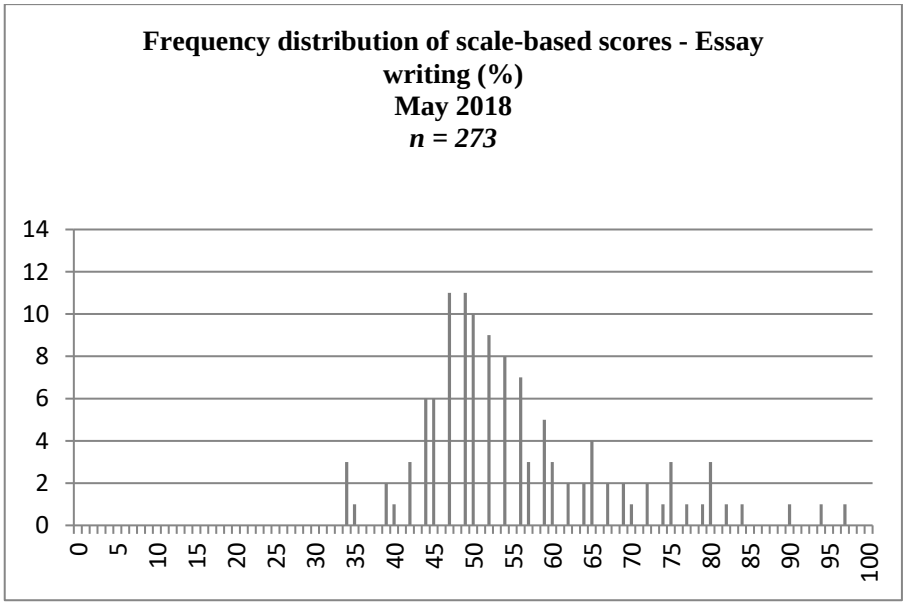


Figure 20. Frequency distribution of reported scores – Essay task

The score distribution is clearly asymmetric and displays a left-skewed curve. The pattern of the score distribution is very similar to that of the July administration (see Figure 12). The central tendency is an indication of the rating scale being incapable of discrimination between low performers and high performers.

When it came to analysing the data for checklist rating, based on the results of Pilot 2, and Research question 2 which was formulated in Chapter 5, I expected (a) increased rater reliability based on the use of dichotomous items, (b) an asymmetric right-skewed distribution instead of the central tendency which was a recurring quality of the score distribution of live administrations. Lastly, I expected (c) greater score variance based on the assumption that the checklist is a better tool for differentiating between low and high performers.

Intra Class Correlation was used to measure the reliability of rating on a checklist between the first and the second rater. For the rater pair A-B the average measure ICC was .906 with a 95% confidence interval from .801 to .955 ($F(29,29) = 10.583$, $p < .001$); for the rater pair B-C the average measure ICC was .800 with a 95% confidence interval from .579 to .0905 ($F(29,29) = 4.988$, $p < .001$); for the rater pair C-D the average measure ICC was .864 with a 95% confidence interval from .783 to .935 ($F(29,29) = 7.360$, $p < .001$); for the rater pair D-A the average measure ICC was .840 with a 95% confidence interval from .654 to .932 ($F(27,27) = 6.239$, $p < .001$). Since average measures above .75 are considered to show high reliability (Educational Authority, 2019a), based on these figures, we may conclude that the reliability of the first and second rating is within the acceptable range. In addition to ICC, I used Krippendorff's alpha to measure inter-rater reliability as well to be able to compare the relationship of double rating using the original accredited C1 Euroexam rating scales with that of checklist-based rating. Based on the raw scores of the double-blind ratings, inter-rater reliability was between $\alpha_K = .621$ and .780 ($p < .05$). Considering the low number of double rated scripts (30), these values are satisfactory. Furthermore, compared to the inter-rater reliability figures of raters using the scale (Table 22), where we observed negative relations, the α_K figures show a considerable increase.

Using item level statistics, I addressed the secondary research questions of Research Question 2, and observed (a) the reliability of the rating tool based on Cronbach's alpha, (b) the facility and quality of individual items, and (c) the ability of the checklist to discriminate low and high performers. The reliability of the checklist was high, $\alpha = .90$. The item level figures are displayed in Table 23.

Table 23
Item-level Statistics of the 30-item Checklist

	SD	*P value*	*Ebel's D*
Item_01	.42	.76	.28
Item_02	.47	.65	.53
Item_03	.47	.33	.78
Item_04	.48	.36	.53
Item_05	.46	.31	.56
Item_07	.43	.75	.13
Item_08	.49	.56	.53
Item_09	.50	.47	.66
Item_10	.50	.49	.84
Item_11	.46	.32	.63
Item_12	.48	.36	.66
Item_13	.48	.37	.59
Item_14	.50	.53	.84
Item_15	.42	.24	.63
Item_16	.48	.64	.38
Item_17	.49	.42	.78
Item_18	.33	.13	.34
Item_19	.46	.31	.75
Item_20	.49	.60	.50
Item_21	.40	.20	.63
Item_22	.42	.23	.44
Item_23	.46	.31	.66
Item_24	.48	.64	.69
Item_25	.41	.21	.50
Item_26	.46	.31	.63
Item_27	.40	.20	.53
Item_28	.50	.47	.59
Item_29	.43	.25	.61
Item_30	.50	.54	.77

The item-level statistics based on the population data show that only one item, Item 1 was easy for the test takers. As discussed earlier, legibility is an important part of writing products therefore, even if it is difficult to formalize and discriminate between low- and high-quality writing, the item was kept to be the part of the checklist. The advantage of checklist-based assessment is that none of the items alone can fail or pass a test taker, this way it gives us more points to consider in the course of assessment. Items 18, 21, 22, 25 and 27 (*The paragraphs create a logically structured text that is easy for the reader to*

follow; This texts demonstrates the use of tenses and aspects; This text uses language to formulate thoughts precisely; The text demonstrates the use of advanced word order and varying sentence length; This text uses a range of discourse functions in a meaningful way.) show low p-values, which indicates that fulfilling these criteria was difficult for the test takers. However, the discrimination indices fall within the acceptable range as defined by the accreditation requirements, so item difficulty on its own indicates that the May 2018 population was a weaker one.

The main aim of the large-sample field test was to compare scale-based and checklist-based rater behaviour and awarded scores. First, I calculated the score distribution for checklist-based scores, then compared the results of the overall judgement using the two different assessment tools, and finally looked at the reliability of the success rate.

The score distribution of the results of checklist-based assessment conformed to my expectations (Figure 21). The bar chart displays a wider and flattened curve compared to score distribution patterns of the scale-based results.

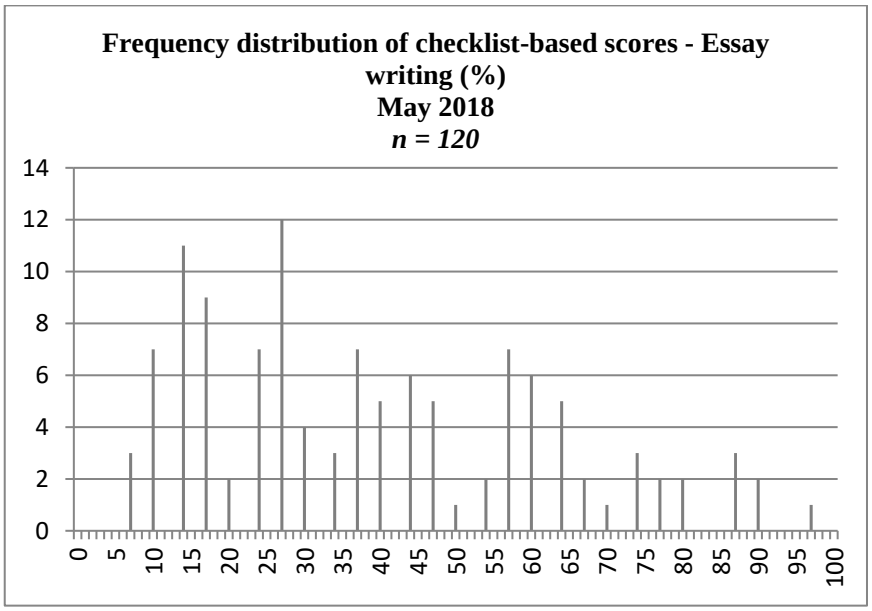


Figure 21. Frequency distribution of checklis-based scores – Essay task

The score variance was much higher than that of the reported scores and on average the scores were lower (M = 38.84; SD = 22.98). The lower mean score and the curve skewed to the right indicate a lower number of high performers, and a higher number of low performers. This, however, does not necessarily mean that the two rating tools differ in terms of severity. In order to reveal whether the two assessment tools mean a more severe rating, or merely a wider score distribution, I compared the success rates for both.

To ground the comparison, first I calculated the inter-rater reliability of first and second checklist-based ratings. There was a high correlation between the overall scores, .846 ($p < .01$) and a very high Krippendorff's alpha, $\alpha_K = .885$. Based on these figures, I considered the first and second ratings equally valid, and in the course of the comparison of scale-based and checklist-based results, I used only the scores of the first raters. The exact agreement between the overall scores of the two rating tools was 79.66%, and the correlation was .703 ($p < .01$). Since one of the aims of the checklist-based rating was to achieve greater score variance, I considered these figures satisfactory.

Another important aspect addressed by the secondary research question of Research Question 2 is concerned with the success rate of the writing paper. In order to reveal whether checklist-based assessment affects success rate, I calculated the success rates for the two rating tools and the computed different statistics to reveal their reliability. The success rate was the same in 86% of the cases, compared to live administration reported scores the checklist provided a proportion of 3.38% false passes and 10.16% false fails. In addition to this, a high degree of reliability was found between the pass-fail rates of checklist-based and scale-based assessment. The average measure ICC was .765 with a 95% confidence interval from .662 to .837 ($F(117,117) = 4.257$, $p < .001$). A paired samples t-test revealed that there was no significant difference in the success rate of checklist-based ($M = .23$, $SD = .422$) and scale-based ($M = .27$, $SD = .446$) rating with the conditions of $t(117) = -1.21$, $p = .227$.

8.9 Conclusion

Based on the findings in Stage 4 of the research-based validation process, it can clearly be stated that the use of the proposed checklist-based assessment tool improves the scoring validity of the essay task of the Euroexam Academic Test. The results of the checklist development process and the large-scale field test support the original research hypothesis. The methodology of the validation research, i.e. mixed-methods research is reflected in the research questions I formulated at the beginning of the research. Research Question 2 targeted the quantitative part of the research, whereas Research Question 3 was used as a qualitative cross-validation. Based on the results of the checklist development project led we may conclude the following:

Research Question 2: Compared with a marking scale, can checklist-based assessment enhance

- the objective scoring of academic discussion essays and
- rater reliability?

As the dichotomous items and the concept check questions leave less chance to rater bias, checklist-based assessment increases rater objectivity. Based on the figures, checklist-based assessment increases the scoring validity of the test. The higher level of inter-rater reliability was demonstrated through various statistical analyses (exact agreement, ICC, and Krippendorff's alpha).

The secondary research questions targeted specific statistics in the course of the development project. Objectively, these figures suggest increased scoring validity for the discussion essay on their own; furthermore, the values fulfil the Hungarian accreditation requirements (Educational Authority, 2019a).

a) Is the reliability (Cronbach's alpha) of checklist scores high enough to fulfil accreditation requirements?
   The reliability of the scores of based on the final 30-item checklist (α = .90) fulfils the accreditation requirement of $\alpha \geq .75$.
b) How do checklist items perform in terms of item difficulty and item quality?
   Concerning item difficulty and item quality the item level statistics conform to the specifications of the *Accreditation Manual*, namely that more than 80% of the p-values and 90% of the discrimination indices (Ebel's D) fall within the acceptable range of .70 ≥ p-value ≥ .30; Ebel's D ≥ .30.
c) Is the checklist capable of discriminating low and high performers?
   The high values for Ebel's D indicate that the checklist can discriminate high and low performers. This is also clearly discernible on the frequency distribution chart (Figure 21), in which we can observe a broader score range without the presence of a central tendency.
d) Does checklist-based rating affect the success rate of the essay task?
   Although the scores are spread out, the success rate of the essay task calculated with a paired samples t-test using scale-based and checklist-based results is not different, therefore checklist-based assessment is not more severe that scale-based assessment, and it does not affect the standard.

As for **Research Question 3** (Can checklist-based marking increase the genre awareness of raters?), the answer may be given based on the results of the qualitative data analysis of Stage 4. The dichotomous items of the checklist and the concept check questions increased

the genre awareness of raters. The majority of the feelings the participants expressed in the course of item development (Table 16) are about increased objectivity and a positive attitude towards a tool that gives clear-cut criteria. They all seemed to be happy to follow these instead of relying on “gut feelings”.

The results of the checklist development project may lead to the conclusion that it is possible to minimise rater bias, reduce the strong central tendency in rating (Eckes et al., 2016), and direct raters toward a common understanding of assessment and genre criteria. Furthermore, an analytical scale that focuses on directly observable phenomena may enhance teacher’s feedback practices and thus increase positive washback.

Chapter 9: Conclusion and Further Research

The aim of this chapter is to give an overview of the dissertation and summarize the main points and findings. After drawing a general conclusion, I highlight the relevance and the implication of the research, touch on its limitations and identify the areas for further research.

9.1 General Conclusion

The dissertation aimed to present the research-based validation process of the writing tasks of the English for Academic Purposes (EAP) test of Euroexam International and the development of a checklist-based rating tool for the assessment of discussion essays within the academic domain. The research project was motivated by the endeavour of Euroexam International to design and implement a locally developed EAP test, and by my interest in the assessment of writing skills and the possible ways of increasing the objectivity of the rating process.

The present study was divided into two main parts: literature review and the stages of my actual research concern. Having set the background to the research in Chapter 1, I reviewed the relevant literature in Chapters 2-4 of the dissertation. In Chapter 5, I presented my methodology and the research questions. The second part, Chapters 6-8, defined the stages for the validation process based on the socio-cognitive framework of Weir (2005a). I identified four stages of the test development and validation research using Read's (2015) approach and deployed qualitative and quantitative methods of investigation. Collecting and analysing quantitative and qualitative data in a research design enabled me to triangulate the research and cross-check findings, and to counteract the distortions associated with the various methods (Creswell et al., 2003). My mixed-methods research was iterative in nature and followed a sequential structure: the results and conclusions of each stage were built in the design of the following stages (Creswell, 2009, p. 14).

Stage 1, the initial development stage started with domain analysis in which apart from reviewing the relevant literature in connection with writing assessment within the academic domain, I also focused on the local context of Euroexam International and the test portfolio they offer. The need for localisation was twofold. On the one hand, I followed O'Sullivan and Dunlea (2015) who propose that that test development projects should always observe the local context and question the relevance of tests that claim to

suit all test taker needs. On the other hand, due to the accreditation requirements, the newly developed test was proposed by Euroexam International as a new profile extension. Conforming to the regulations of the Accreditation Board and the Euroexam portfolio meant that the development process of the writing tasks that were specifically designed for the academic test had to observe the construct definition of the Euroexam writing tasks. This way, the domain analysis stage investigated whether and to what extent the existing writing task types (Task 1: *transactional writing* and Task 2: *discursive writing*) could be valid measures of writing skills within the academic domain.

To generate validity evidence for the context validity of transactional writing in the academic domain, I carried out a preliminary study based on university student (N = 5) and staff (N = 6) interviews. I investigated what students write and how they communicate with university staff and aimed at establishing the validity of transactional writing as part of the academic discourse. Based on the answers to my secondary research questions of the preliminary investigation, my original hypothesis was confirmed: transactional writing is part of the academic domain. The mixed-methods research design was iterative in nature and involved triangulation to cross validate the findings of the elements of Stage 1 of the research. The results of the preliminary investigation were used to define the construct of the new Euroexam Academic test. The construct definition and the preliminary task design were the subject of the external expert judgement (N = 3). To enhance validity, the external experts were not provided with the empirical findings, but they reviewed the construct and the example tasks using a questionnaire. Using their knowledge and experience already available in the field, the experts found the proposed test tasks valid representations of the academic domain, and their comments helped me draft the specifications.

As for the validity evidence of discursive writing in the domain, I draw on literature review and used expert judgement to see what genres could be used as valid tasks in an EAP test. Although numerous written genres appear as typical in academic tests, such as essay, report, summary, library research paper (Carson, 2001; Cooper & Bikowski, 2007; Hale et al., 1996), based on the test review by the invited external experts, I decided to use only the genre of essay in Task 2 in three distinct fields of study. This way, the three options that appear in Task 2 may be chosen based on test taker preference, while the use of one specific genre makes test takers' results comparable in different test administrations. By using transactional writing and discursive writing for Task 1 and Task 2, respectively, it

was possible to keep the writing construct as specified by the requirements of profile extension in the *Accreditation Manual* (Educational Authority, 2019a).

After Stage 1, the aim of Stage 2 was to complete the test specifications and the example tasks that were to be pretested in Stage 3. Similarly to domain analysis, domain modelling in Stage 2 is also evidence focused. I used test taker performance and test taker and rater interviews to see whether the two proposed task types conform to the construct. At this point, it is important to highlight that the verbal protocols and the textual analysis in Stage 2 also served as the trialling of the exam tasks. When test taker and rater feedback was collected, I used the preliminary tasks which were designed at the end of Stage 1 for think aloud and immediate recall. The reason for this is twofold. On the one hand, the theoretical construct of the genres is not suitable for collecting user feedback. On the other hand, through trialling example tasks, we may observe test taker characteristics and task characteristics; in other words, the validity of the rating process may be ensured.

Chapter 7 reports on domain modelling, the trialling and pretesting of the proposed test tasks in Stage 2 and Stage 3. As mentioned above, domain modelling also involved trialling the test tasks and served the purpose of mapping the skills test takers utilise when completing the tasks (Perie & Huff, 2016). The small-sample trial was conducted using qualitative methods: potential test taker (N = 6) interviews and Euroexam rater think-aloud protocols. The main aim of Stage 2 was to complete the test specifications, which was done based on the data gathered in the course of trialling. I gathered data on the validity of the rating process using experienced accredited raters (N = 3) of Euroexam International in Stage 2. The three raters were given the same test taker scripts and were asked to assess them when using the accredited C1 level writing scale of Euroexam International through a think-aloud technique in Hungarian. I found that Euroexam raters varied in their scoring behaviour, their construct interpretations, and their severity. It seemed that the vague descriptors of the rating scale hindered the objectivity of the rating process and thus the reliability of the test scores (see the results in Table 8, Table 11 and the accredited rating scale in Appendix 6).

In accordance with the cyclical nature of the design, at the end of Stage 2, I redesigned Task 1 based on the students' comments. In addition to this, a detailed task specification was completed for the Euroexam Academic Test. In Stage 3, the large-scale pretesting stage, I used the format and layout of the tasks as they were redesigned based on Stage 2.

The validation process in Stage 3 involved large-scale data collection and evidence-based analysis of test taker performance. The aim of this stage was to check that test tasks work as intended so that the standard level of the Euroexam Academic test (C1) could be set, the relationship to the CEFR could be established and the validity of the test could be demonstrated. To ensure all this, I used a sample size (N = 136) for pretesting that allows statistical data analysis using Classical Test Theory (CTT). The aim of pretesting is to model the live administration of the test and to see how test takers and test tasks perform under exam circumstances. Together with pretesting the tasks of the Academic Test, I used a questionnaire I designed to collect test taker personal data concerning their language learning background and self-assessment as well as test taker opinion of the form and content of the test.

The statistical analysis of test taker results (see Table 9-11) proved that the Euroexam Academic Test is a valid measure of C1 level writing skills, however, raised further issues about the scoring validity of the two writing tasks. In addition to the Writing results, I also analysed the results on the Reading and Listening papers as these papers always contain common items, in other words repeated tasks, to make sure that the different test batteries are comparable across administrations. I calculated inter-rater agreement and observed correlations between the results on the different test papers (see p. 114). I used the results on the Listening and Reading papers and the correlation between the repeated tasks and the Academic tasks to cross validate the results of my analysis of the subjectively marked Writing tasks (see Table 11). The repeated tasks seemed to have been easier whereas the Academic tasks were more challenging for the population – still the pretest population was on average 12% more able than the live administration. The higher ability of the test takers, however, was not reflected in the writing results. It was rather unusual to see that the writing results were the closest to the cut-off score; and there was no difference between the General C1 and the Academic population. I interpreted this result as indication of rater behaviour, especially because the correlation and the inter-rater agreement between the 1st and 2nd rater was relatively low.

I administered student questionnaires (see Appendix 9) that asked about language learning background, self-evaluation, and the evaluation of the test papers. The questionnaire revealed test takers' perception of the different tasks and their own performance which could be compared regarding their overall results and their writing results. I performed a chi square test of independence which showed that there is strong

significant relationship between the test takers' predictions concerning their own general performance and their results. Interestingly, this perception is completely different for the writing tasks (see p. 114). The results of the test show that there is no significant association between students' perception and the results of their own writing performance. We may conclude that this is because of a strong rater effect, perhaps together with the test takers' vague ideas about a high-quality writing product.

Due to the recursive process in my data analysis, it became clear that the results of the earlier stages all lead in a certain direction: they highlighted raters' differences and flaws of the rating procedure. The results of Stage 2 and Stage 3, based on the verbal protocols, revealed that there was a discrepancy between the scores and the students' self-evaluation. The shortcomings of the C1 level accredited rating scale of Euroexam International were also exposed by the rater think aloud protocols. Chapter 7 revealed raters' ideas about the writing product and the rating scale and revealed considerable rater bias. Therefore, I added an additional research stage to the standard stages of validation with the aim of designing a more objective tool for the assessment of the discussion essay task that compensates for individual rater characteristics and rater effect.

Stage 4 of the research focused on the development of a checklist-based assessment tool following Lukácsi's (2017; 2018; 2020) research to increase the scoring validity and the reliability of the assessment of the essay task. The aim of this stage was to develop a task and level-specific checklist-based rating tool for the essay writing task. As it is visible in Table 3, Stage 4 consists of two main parts, document analysis (phase 1) and empirical research (phases 2-8). The empirical research was divided into two major steps: phases 2-6 focused on designing and developing the items of the checklist, and phases 7-8 aimed at exploring the relationship between scale-based and checklist-based scores.

Since test fairness and validity are closely connected, it was important to foster the common understanding of the rating tool and through that the writing construct. The Stage 4 research project included the review of literature in connection with the suitability of different assessment tools for writing tasks together with qualitative and quantitative data analysis. Data collection was based on verbal protocols and the analysis of test taker performance using Classical Test Theory (CTT). Based on the mixed methods of the 8 phases, I developed a 30-item checklist with dichotomous items (Appendix 16), together with concept check questions to enhance the construct interpretation of raters. The new rating tool had to undergo a number of statistical procedures in order to make sure that the

figures concerning its quality and reliability conform to the requirements of the *Accreditation Manual*. Based on calculations using the tools of CTT, I found that he success rate was the same in 86% of the cases (see p. 152). Also, a paired samples t-test revealed that there was no significant difference in the success rate of checklist-based and scale-based rating. In addition to this, a high degree of reliability was found between the pass-fail rates of checklist-based and scale-based assessment. The analyses confirmed that the level and genre specific checklist enhances the objectivity and reliability including the following advances: (a) broader score range, (b) increased accountability and (c) transparency. In general, the checklist as a rating tool has been proved to be more fitting for level testing.

The similar score distributions in different parts of the test may lead to a number of conclusions. It became clear that checklist-based assessment reduces the bad practice of central tendency (see Figure 18 and 19; Figure 20 and 21), and the score distribution will mirror the score distribution in the Reading part of the test (see Figure 13 and 19). With the checklist-based rating tool, it was possible to reduce rater bias and achieve a fairer rating, which are reflected in the similar score distribution of the objective, machine rated and the human rated scores. Furthermore, the broader score range concerns the two approaches, the compensatory and the conjuctive approach (Government Decree 137/2008 (V. 16.) that are present simultaneously in Hungarian accredited language testing. According to the compensatory approach, failing on some tasks may be compensated by good performance on other tasks; whereas, according to the conjunctive approach, each skill is tested and assessed separately. Test takers need to achieve at least 60% overall to pass the level test, but a 40% minimum performance counts as a compensable fail, which means that failing one task can be compensated by better performance on a different task. The score distribution displayed in Figure 20 highlights that applying the compensatory approach will pass almost all the test with a better result on the Reading paper. This leads us to the conclusion that increased variation also increases scoring validity, reduces rater effect, and is more suited to the special requirements of the local Hungarian accredited language assessment system.

Based on the results of the checklist development project, we can claim that the checklist-based rating tool has a number of advantages. All things considered, the research design could be adapted to develop a similar rating tool for the transactional writing task,

as well as all for the genres that appear in the writing paper of the C1 level Euroexam General English Test.

9.2 Implications

The most important findings of the dissertation concern the validity of the writing tasks of the locally developed EAP test of Euroexam International. Based on the results of the 4-stage research-based validation process, it has been confirmed that the test tasks are valid measures of English language skills within the academic domain.

A further contribution of the research is the development and validation of a checklist-based rating tool, the use of which results in an increased scoring validity and a more reliable rating for the discussion essay task. Apart from increasing the statistical reliability of rating, the additional aim of Euroexam to strengthen raters' awareness for good quality essays, increase objectivity and compensate for harshness/leniency seems to be fulfilled. The advantage of the checklist lies in the use of dichotomous items, additional concept check questions and transparent instructions. Furthermore, the items help maintain the construct relevant nature of the assessment. The interviews and the think-aloud protocols supported the validity of the results of the statistical analysis.

Another major implication for large-scale high-stakes testing concerns rater training. Increasing the objectivity of the rating procedure and reducing the compulsory training and re-training hours is in the interest of exam providers in general. All the more so because the potential benefits of rater training, as pointed out in Chapter 4, do not outweigh the time and effort spent on it in each test administration session (Weigle, 1994). The checklist-based rating tool together with the concept check questions that are added to the dichotomous items, proved to be a reliable assessment tool that enhances raters' unbiased decisions.

As for face validity issues, the effect of the checklist is expected to be an increased validity in case of test takers. Transparency or relevance of a test becomes visible for test takers through the rating and the test results. Since the concept behind the items is defined with very clear criteria using the CEFR scale descriptors, and the statements of the checklist are complemented with concept check questions, rater accountability will increase. In case of a possible test taker appeal process after failing the test, it would be easier for raters to highlight and point out the shortcomings of test taker performance, and the results would appear more transparent and thus more acceptable to test takers.

Another point to consider as an important implication of my research concerns rater fatigue. The interview in the development phase covered this area and raters were invited to express their feelings in connection with using a four-page long complex rating tool. The users, however, did not complain about an extremely long rating time. They reported 5-15 minutes for a script, and they also reported that they learned the order of the items very quickly, which meant that first it took longer to assess a writing product, then they became faster, and after a certain point they slowed down again, which is definitely a sign of the so called "judgemental fatigue" (Council of Europe, 2001, p. 50). It has been revealed that fatigue may introduce construct-irrelevant factors to a test score and affect its validity and fairness, but it may be compensated by shorter rating sessions (Ling et al., 2014). Limiting the number of scripts per rater and shortening the rating time a day can maintain greater rating productivity, accuracy, and consistency. However, these are factors beyond the validity of the rating device itself. Another advantage of the checklist is the use of precise statements and concept check questions, which may slow down raters after a certain point, but do not let them bring in construct irrelevant factors.

It is important to point out here that the studies in connection of rater fatigue mainly concern rating scales, which means that scales and checklists operate in a similar way in this respect. The solution to the problem may lie in maximising number of writing products the raters assess. In the checklist development project, the maximum number of scripts allocated to participants was 30, which did not lead to extreme signs of judgemental fatigue. Ling et al. (2014) also idealize shorter shifts and introducing breaks, which means that based on professional considerations, it would be acceptable to allocate a maximum of 50 writing products per person per day to avoid validity and fairness issues.

One other advantage of a more objective rating may be identified regarding the washback effect of the test. Providing the stakeholders of the Academic test (teachers, instructors, future test takers) with a clearly defined set of criteria may result in improving the writing skills of test takers (Wall, 2005) and thus increase their success rate at the Euroexam Academic Test. In the literature review (Chapter 4.2), I summarise the main points of the relevant literature in connection with the washback effect, considering the negative and positive washback as well. According to testing literature, positive washback occurs when teachers use classroom practices that enhance language ability and promote test preparation at the same time. The binary choice statements of the checklist I propose are highly suitable for classroom-based preparation as the cover specific areas of language

ability at a given level. Instead of the vague descriptors of the scale, teachers will have tangible prompts for practice, test preparation, and diagnostic assessment. In this regard, my contribution is highly relevant for English instructors due to the positive impact and the washback effect the rating process might have on exam preparation courses.

9.3 Limitations and further research

As pointed out in Chapter 5 and Chapter 6, the validation research may be interpreted within the context of the Hungarian accreditation system for foreign language exams (Educational Authority, 2019a) and the C1 level writing construct of the tests of Euroexam International (Euroexam International, 2019b). This contextualisation might be understood as a limitation of the research; however the dissertation acknowledged this framing, and argued for the necessity of localisation as proposed O'Sullivan & Dunlea, 2015.

The pretesting stage described in Chapter 7 considered how the score distributions for the three skills in the written papers relate to each other, and also how the Academic pretest students performed at the academic tasks and the repeated tasks. In Chapter 8, I looked at if and how the checklist-based writing scores mirrored performance on other skills, but after a number of live administration sessions, it would be beneficial to design further research projects to see how Euroexam Academic test takers perform regarding all four skills.

In the present research, textual analysis of test taker performance was only a small-scale endeavour. In the future, it would be advantageous to look at the relationship between raters' assessment and the textual and linguistic features of writing products using automated text analysis tools. The use of consistent text analysis categories of these tools would make large-scale analysis easier, more consistent, and comparable across different administrations.

Since I used Classical Test Theory for item analysis, I had to be aware of the population dependent nature of the results. Despite its shortcomings, CTT is widely applied in the field. In my research project, I attempted to counteract these drawbacks and cross validate the findings by using mixed-methods research and triangulation. It is plausible that the limitations of CTT may have influenced the results. Consequently, further analysis using modern test theory would be useful in the future to analyse the performance of the rating tool and the performance of test takers of the Euroexam Academic test.

All things considered, based on the results of the checklist development project (Chapter 8), I can claim that the checklist-based rating tool has a number of advantages over the accredited rating scale. Therefore, the research design should be adapted to develop a similar rating tool for the transactional writing task as well as all the genres that appear in the writing paper of the C1 level Euroexam General English Test (report, review, online comment, online article). As the checklist-based rating tool is genre and level specific, I expect my research to contribute to the future development projects of Euroexam International regarding the development of more objective and reliable rating tools for the assessment of subjectively assessed test tasks at each CEFR level that appears in their test portfolio. As part of the continuation of my current project, further research should be conducted to examine how raters approach the assessment of speaking tasks to see how a checklist-based rating tool might be designed and implemented for the assessment of oral proficiency.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C., & Clapham, C. (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly, 29*(1), 184–187.

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

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. *Language Testing, 22*, 301–320.

ALTE (2011). *Manual for language test development and examining*. Strasbourg: Council of Europe.

Anastasi, A. (1988). *Psychological testing* (6th edition). New York: Macmillan.

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

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

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F. (2008). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149–64.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Banerjee, J., & Harsch, C. (2016). *Rating the construct reliably*. Presentation at the EALTA Summer School, Innsbruck, 2016. Retrieved on 16 June 2020 from http://www.ealta.eu.org/events/Summer_school_2016/07_EALTA%20SuSch_2016_Rater%20training_part1.pdf

Bárdos, J. (2002). *Az idegen nyelvi mérés és értékelés elmélete és gyakorlata*. Budapest: Nemzeti Tankönyvkiadó.

Bárdos, J. (2003). A nyelvtudás megítélésének korlátai. *Iskolakultúra*, *13*(8), 28–39.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75.

Benke, E. (2007). *An investigation of rater and rating scale interaction in the validation of the assessment of writing performance*. PhD thesis. Eötvös Loránd Tudományegyetem, Budapest.

Bereiter, C., & Scardamalia, M. (1982). From conversation to composition: The role of instruction in a developmental process. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol 2, pp. 1–64.). Hillsdale, NJ: Erlbaum.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brindley, G. (2000). Task difficulty and task generalizability in competency-based writing assessment. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (Vol. 1, pp.125–157). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Bukta, K. (2007). *Processes and outcomes in L2 English written performance assessment: Raters' decision-making processes and awarded scores in rating Hungarian EFL learners' compositions*. PhD thesis. Pécsi Tudományegyetem, Pécs.

Bukta, K. (2013). *Rating EFL written performance*. London: Versita. https://doi.org/10.2478/9788376560793

Butler, F. A , Weigle, S. C., Kahn, A. B., & Sato, E. Y. (1996). *Test development plan with specifications for placement instruments anchored to the model of standards.* Los Angeles: University of California, Los Angeles.

Butler, J. (1990). *Gender trouble: feminism and the subversion of identity*. New York and London: Routledge.

Byram, M., & Parmenter, L. (Eds.) (2012). *The common European framework of reference: The globalisation of language education policy*. Bristol, UK: Multilingual Matters.

Byrnes, H. (Ed.) (2006). *Advanced language learning: The contribution of Halliday and and Vygotsky*. London–New York: Continuum

Cameron, D. (2000). Styling the worker: Gender and the commodification of language in the globalized service economy. *Journal of Sociolinguistics*, *4*(3), 323–347.

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher, & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections*, (pp. 48–83). Ann Arbor, MI: The University of Michigan Press.

Carter, M. (2007). Ways of knowing, doing, and writing in the disciplines. *College Composition and Communication, 58*(3), 385–418.

Chan, S. H. C. (2013). *Establishing the validity of reading into writing test tasks for the UK academic context*. PhD thesis. University of Bedfordshire. Retrieved on 16 June 2020, from: http://uobrep.openrepository.com/uobrep/handle/10547/312629

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Grabe, W., & Berns, M. (1993). *Communicative language proficiency: Definitions and implications for TOEFL 2000.* ETS Internal Report. Princeton, NJ: Educational Testing Services.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Collado, a. V. (1981). Using students' first language: Comparing and contrasting. *TESOL Higher Education Interest Section Newsletter 3*(9), 9–10.

Connor, U. (1997). Contrastive rhetoric: Implications for teachers of writing in multicultural classrooms. In C. Severino, J. Guerra and J. Butler (Eds.), *Writing in*

multicultural settings (pp. 198–208). New York: Modern Language Association of America.

Connor, U. (2004). Intercultural rhetoric research: beyond texts. *Journal of English for Academic Purposes, 3*(4), 291–304.

Cooper, A., & Bikowski, D. (2007). Writing at the graduate level: what tasks do professors actually require? *Journal of English for Academic Purposes, 6,* 206–221.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved on 20 June 2020, from https://rm.coe.int/1680459f97

Council of Europe (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR). A manual.* Strasbourg: Language Policy Division. Retrieved on 20 June 2020, from https://rm.coe.int/1680667a2d

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Language Policy Division. Retrieved on 20 June 2020, from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Creswell, J.W. (2009). *Research design. Qualitative, quantitative and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*(2-3), 170–191. https://doi.org/10.1504/IJCEELL.2011.040197

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*, 1–8.

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Vol. Ed.), *The companion to language assessment: vol. 1*, (pp. 216–229). Hoboken: John Wiley & Sons, Inc.

Cumming, A, Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5–43.

Cumming, J., & Maxwell, G. (1999). Contextualising authentic assessment. *Assessment in Education*, *6*, 177–194.

Davidson, F. and Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Davies, A. (2007). Assessing academic English language proficiency: 40+ years of U.K. language tests. In J. Fox et al. (Eds) *Language testing reconsidered.* (pp. 73-88) Ottawa: University of Ottawa Press.

Davies, A. (2003). Three heresies of language testing research. *Language Testing 20*(4), 355–68.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521–541.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2017). University entrance language tests: A matter of justice. *Language Testing*, *35*(4), 449–476. https://doi.org/10.1177/0265532217706196

Di Gennaro, K. (2006). Second language writing ability: Towards a complete construct definition. *Working Papers in TESOL and Applied Linguistics*, *6*(2), 1–17.

Dörnyei, Z. (1988). Language testing. In Z. Dörnyei, *Psycholinguistic factors in foreign language learning* (pp. 8–46). PhD thesis. Budapest: Eötvös Loránd University.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185.

Eckes T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the common European framework of reference for languages*: *Learning, teaching, assessment (Section H)* (pp. 1–52). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In: D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 147–164). Boston: De Gruyter Mouton.

Educational Authority (2019a). *Accreditation manual 2019*. Retrieved on 16 June 2020, from https://nyak.oh.gov.hu/nyat/doc/ak2019/ak2019.htm

Educational Authority (2019b). *Language test statistics (Based on the number of test takers)*. Retrieved on 14 June, 2020 from https://nyak.oh.gov.hu/doc/statisztika.asp

Educational Authority (2020a). *Accredited language tests*. Retrieved on 14 June, 2020 from: https://nyak.oh.gov.hu/doc/akk_vizsgarendszer.asp.

Educational Authority (2020b). *Basic definitions*. Retrieved on 16 June 2020, from: https://nyak.oh.gov.hu/doc/alapfogalmak-eng.asp

Educational Authority (2020c). *Nationalisation*. Retrieved on 16 June 2020, from: https://nyak.oh.gov.hu/doc/honositas/honositas_nyelv.asp

Educational Testing Services (2019). *TOEFL iBT® Test independent writing rubrics*. Retrieved on 16 July 2020, from: https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

Eduline, (2017, October 4). *Itt a 2018-as középiskolai rangsor: ez a tíz legjobb budapesti gimnázium*. Retrieved on 19 June 2020, from https://eduline.hu/kozoktatas/legjobb_budapesti_gimnaziumok_rangsor_7IB97R

Elder, C. (2014). Book review: The Routledge handbook of language testing. *Language Testing 31*(1), 138–144.

Elder, C., Knoch, U., Barkhuizen, G., & Randow von J. (2005). Individual feedback to enhance rater training: does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research, 4*(3), 193–220. https://doi.org/10.1177/136216880000400302

Endres, H. (2012). A comparability study of computer-based and paper-based writing tests. *Research Notes, 49*, 26–33.

English Profile (2015). *English vocabulary profile online*. Retrieved on 16 June 2020, from: https://www.englishprofile.org/wordlists/evp

Erdosi, M.U. (2001). The influence of prior experience on the construction of scoring criteria for ESL compositions: A case study. *International Journal of English Studies 1*(2), 175–196.

Euroexam International (2018a). *C1 level General English results – Live administration May 2018*. Raw data: unpublished.

Euroexam International (2018b). *C1 level General English results – Live administration July 2018*. Raw data: unpublished.

Euroexam International (2018c). *Guide for item writers*. Internal Document. Budapest: Euroexam International.

Euroexam International (2019a). *Euroexam Academic English C1*. Retrieved on 17 June 2020, from: http://www.euroexam.com/the-exams/euroexam-academic

Euroexam International (2019b). *Euroexam detailed specifications.* Retrieved on 16 June 2020, from https://rex.oh.gov.hu/FileDb/0015-AAAAAH/specifikacio/190206-003_EURO.doc

Euroexam International (2020). *Euroexam level C1*. Retrieved on 16 June 2020, from http://www.euroexam.com/sites/network/files/file/download/Marking_Criteria/EuroC1WritingScalesLevelDescriptorsE.pdf

Extra, G., Spotti, M., & van Avermaet, P. (2009): Testing regimes for newcomers. In: G. Extra, M. Spotti, & p. Van Avermaet. (Eds.), *Language testing, migration and citizenship* (pp. 3–33). New York, NY: Continuum.

Fairclough, N. (1999). Global capitalism and critical awareness of language. *Language Awareness, 8*(2), 71–83.

Fairclough, N. (2001). *Language and power*. London: Longman.

Ferris, D. R., & Hedgecock, J. S. (2005). *Teaching L2 composition: Purpose, process, and practice*. New York: Routledge.

Field, J. (2004). *Psycholinguistics: The key concepts.* London: Routledge.

Figueras, N., & Noijons, J. (Eds). (2009). *Linking to the CEFR levels: Research perspectives.* Arnhem: CITO-EALTA. Retrieved on 16 June 2020, from: http://www.ealta.eu.org/documents/resources/Research_Colloquium_report.pdf

Flower, L., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. Gregg and E. R. Sternberg (Eds.), *Cognitive processes in writing* (pp. 31–50). Hillsdale: Erlbaum.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research in writing: Principles and methods* (pp. 75–98). New York: Longman.

Freimuth, H. (2017). Revisiting the suitability of the IELTS examination as a gatekeeper for university entrance in the UAE. In: L. Buckingham (Ed.), *Language, identity and education on the Arabian Peninsula: bilingual policies in a multilingual context* (pp.161-175). Bristol: Multilingual Matters.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics 20*(2), 221–236.

Fulcher, G. (2010). *Practical language testing*. New York: Routledge.

Fulcher, G. (2016). Standards and frameworks. In: D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 29–44). Boston: De Gruyter Mouton.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing 26*(1) 123–144.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Fűköh, B. (2016). Developing Writing Skills of Students of Business English. *NyelvVilág, XX,* 7–18.

Fűköh, B. (2018). Student interviews in establishing the context validity of an EAP writing task. In: (Besznyák R. (Ed.) *Porta Lingua - 2018. Tudásmegosztás, értékközvetítés, digitalizáció– trendek a szaknyelvoktatásban és -kutatásban* (pp. 301–309). Budapest: Szaknyelvoktatók és -Kutatók Országos Egyesülete. http://szokoe.hu/kiadvanyok/porta-lingua-2018

Fűköh, B. (2019a, May). *Research-based EAP test development: local needs and opportunities on an international context.* Paper presented at the 14th UPRT Empirical Studies in Applied Linguistics Conference, University of Pécs, Pécs, Hungary.

Fűköh, B. (2019b). Kutatáson alapuló tesztfejlesztés – írásfeladatok egy angol tudományos szaknyelvi vizsga számára. In: (R. Besznyák (Ed.) *Porta Lingua – 2019. Interdiszciplináris megközelítések a szaknyelvoktatásban és –kutatásban* (pp. 271–288). Budapest: Szaknyelvoktatók és -Kutatók Országos Egyesülete. http://szokoe.hu/kiadvanyok/porta-lingua-2019

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. London: Lawrence Erlbaum Associates, Publishers.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing 26*(4), 507–531.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40*, 97–118.

Given, L. M. (2008). *The SAGE encyclopaedia of qualitative research methods*. Thousand Oaks, CA: SAGE Publications Ltd.

Government Decree 137/2008. (V. 16.) Korm. rendelet az idegennyelvtudást igazoló államilag elismert nyelvvizsgáztatás rendjéről és nyelvvizsga bizonyítványokról. Retrieved on 16 June 2020, from: http://njt.hu/cgi_bin/njt_doc.cgi?docid=119127

Grabe, W., & Kaplan, R.B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman, New York.

Grabowski, J. (2005). Speaking, writing, and memory span performance: Replicating the Bourdin and Fayol results on cognitive load in German children and adults. In L. Allal, & J. Dolz (Eds.), *Proceedings Writing 2004*. Geneva (CH): Adcom Productions. [CD-ROM]

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2011). Methods of Automated Text Analysis. In: M. L. Kamil, D. P. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (pp. 34–53). New York: Taylor & Francis.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree program*. *TOEFL Research Reports*. RR-95-44. Princeton, NJ: Educational Testing Service.

Hall, G. (2010). International English language testing: A critical response. *ELT Journal 64*(3), 321–328.

Halliday, M. A. K. (1994). *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. (2004). *An introduction to functional grammar*. London Routledge.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*, 759–762.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In: Barbara Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge, UK: Cambridge University Press.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community, and assessment*. (TOEFL Monograph Series report No. 5). Princeton, NJ: Educational Testing Service.

Harsch, C. (2018). How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly, 15*(1), 102–108.

Harsch, C., & Banerjee, J. *The construct of writing.* Presentation at the EALTA Summer School, Innsbruck, 2016. Retrieved on 16 June 2020 from http://www.ealta.eu.org/events/Summer_school_2016/02_EALTA%20SuSch%202016_Construct%20of%20writing.pdf

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228–250. https://doi.org/10.1016/j.asw.2012.06.003

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education*, *20*(3), 281–307.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33. https://doi.org/10.1080/15434303.2010.535575

Harsch, C., & Seyferth, S. (2020). Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist. *Assessing Writing, 43,* 1–15. https://doi.org/10.1016/j.asw.2019.100433

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT® test score and their use in informing policy in a United Kingdom university setting.* (ToeFL iBT Research Report No. 30). Princeton, NJ: Educational Testing Service. https//doi.org/10.1002/ets2.12167

Hartley, J. (2008). *Academic writing and publishing*. London: Routledge.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York, NY: Newbury House Publishers.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy, & S. Ransdell (Eds.), *The Science of writing* (pp. 1–27). NJ: Lawrence Erlbaum Associates.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg, & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum.

Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction*. Ann Arbor, MI: University of Michigan Press.

Hocks, M. (2003). Understanding visual rhetoric in digital writing environments. *College Composition and Communication, 54*(4), 629–656.

Hughes, A. (Ed.) (1988). *Testing English for university study: ELT Documents 127*. Oxford, UK: Modern English Press.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hyland, K. (2002). *Teaching and researching writing*. London: Longman.

Hyland, K. (2004). *Second language writing*. Cambridge, UK: Cambridge University Press.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for Academic Purposes. 1*(1), 1–12.

Hyland, K., & Jiang, F. K. (2017). Is academic writing becoming more informal? *English for Specific Purposes 45*, 40–51. https://doi.org/10.1016/j.esp.2016.09.001

IELTS (2018). *IELTS test format.* Retrieved on 16 June 2020, from: https://www.ielts.org/about-the-test/test-format

IELTS (2020). *IELTS writing mark schemes*. Retrieved on 16 June 2020, from https://www.examenglish.com/IELTS/IELTS_Writing_MarkSchemes.html

Inoue, Ch., Kabbazbashi, N., Lam, D., & Nakatsuhara, F. (2018, November) *The IELTS speaking test: What can we learn from examiner voices?* Paper presented at the Language Testing Forum, CRELLA, University of Bedfordshire, UK. Retrieved on 14 June, 2020 from: https://ukalta.org/wp-content/uploads/2017/10/Inoue-et-al-LTF2018_1.pdf

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics*, *34*, 123–153.

Kane, M. T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational measurement*. (4th ed.). (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds). *Handbook of test development* (pp. 64–80). New York and London: Routledge.

Kane, M. T. (2019, May). *Flexibility and utility in assessment and validation*. Paper presented at the 16th EALTA Conference, University College of Dublin, Dublin, Ireland.

Kaplan, R. (1966). Cultural thought patterns in intercultural education. *Language Learning 16*(1), 1–20.

Kellog, R. T. (1994). *The psychology of writing.* New York: Oxford University Press.

Kim, Y-H. (2011). Diagnosing EAP writing ability using the Reduced Reparametrized Unified Model. *Language Testing*, *28*(4), 509–541. https://doi.org/10.1177/0265532211400860

Kiszely, Z. (2003). *Students' writings in L1 Hungarian and L2 English: Rhetorical patterns, writing processes and literacy backgrounds*. PhD Thesis. Pécsi Tudományegyetem, Pécs.

Kiszely, Z. (2006). Magyar és angol nyelvű fogalmazások retorikai szerkezete: összefüggések, magyarázatok és pedagógiai implikációk. *Magyar Pedagógia, 206*(2), 129–146.

Knoch, U. (2009). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*, 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments. *Papers in Language Testing and Assessment, 2*(2), 1–19.

Knoch, U., Rouhshad, A., Ping O. S., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28*, 39–52.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second Language Writing: Research insights for the classroom* (pp. 37–57). Cambridge: Cambridge University Press.

Krashen (1985). *The input hypothesis*. London: Longman.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Lado, R. (1961). *Language testing*. London: Longman.

Land, R. E., & Whitley, C. (1989). Evaluating second language essays in regular composition classes: Toward a pluralistic U.S. rhetoric. In D. M. Johnson, & D. H. Roen (Eds.), *Richness in writing: Empowering ESL students* (pp. 284–293). London and New York: Longman.

Lane, S, Raymond M. R., & Haladyna, T. M. (Eds.) (2016). *Handbook of test development*. New York and London: Routledge.

Lantolf, J. P. (2000). Introducing sociocultural theory. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 1–26). Oxford: Oxford University Press.

Lazaraton, A., & Taylor, L. (2007). Qualitative research methods in language test development and validation. In J. Fox et al. (Eds.) *Language testing reconsidered* (pp. 113–129). Ottawa: University of Ottawa Press.

Leedham, M. (2015) *Chinese students' writing in English: Implications from a corpus-driven study*. Oxford: Routledge.

Leki, I. (1992). *Understanding ESL writers.* NH: Heinemann Educational Books.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*(1), 43–64.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing 31*(4), 479–499. https://doi.org/10.1177/0265532214530699

Lukácsi, Z. (2013). *Cohesion and writing quality: exploring the construct of cohesion in Euro Examinations*. PhD Thesis. Pécsi Tudományegyetem, Pécs.

Lukácsi, Z. (2017, May). *Developing a level-specific checklist for assessing writing*. Paper presented at the 14th EALTA Conference, Ciep, Sèvres, France.

Lukácsi, Z. (2018). Írásművek lista alapú értékelése – avagy hogyan mérjünk a nyelvvizsgán. *NyelvVilág, XXI,* 7–23.

Lukácsi, Z. (2019a). *Euroexam Academic pretest report: Reading and Listening Papers* Internal Report: unpublished.

Lukácsi, Z. (2019b). Új utak a nyelvi mérésben. *Modern Nyelvoktatás, 25*(3-4), 45–74.

Lukácsi, Z. (2020). Developing a level-specific checklist for assessing EFL writing. *Language Testing*. https://doi.org/10.1177/0265532220916703

Lukácsi, Z., & Fűköh, B. (2018). *Assessing academic English among Central European students for UK University admissions*. Poster presented at the Language Testing Forum 2018 Conference, University of Bedfordshire, Luton, UK.

Lukácsi, Z., & Fűköh, B. (2019, May). *Assessing academic English with a localised test for international admissions purposes*. Paper presented at the 16th EALTA Conference, SIG Assessment of Writing/Assessment for Academic Purposes, University of Dublin, Ireland.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lynch, B. K., & Davidson, F. (1994). Criterion-referenced language test development: Linking curricula, teachers and tests. *TESOL Quarterly*, *28*(4), 727–743.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Matalene, C. (1985). Contrastive rhetoric: An American writing teacher in China. *College English 47*, 789–807.

Matsuda, P. K. (1997). Contrastive rhetoric in context: A dynamic model of L2 writing. *Journal of Second Language Writing*, *6*(1), 45–60.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, *18*, 333–349. https://doi.org/10.1191/026553201682430076

McNamara, T. (2012). Language assessments as Shibboleths: A poststructuralist perspective. *Applied Linguistics*, *33*(5), 564–581. https://doi.org/10.1093/applin/ams052

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

McNamara, T. F. (2000). *Language testing*. Oxford, UK: Oxford University Press.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576.

Menken K. (2017). High-stakes tests as de facto language education policies. In: E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment. Encyclopaedia of language and education* (3rd ed.) (pp. 385–396). Springer, Cham. https://doi.org/10.1007/978-3-319-02261-1_25

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Research Report 90–11. Education Testing Service.

Messick, S. (1994). *Alternative modes of assessment, uniform standards of validity*. Princeton, N.J.: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01634.x

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher 23*(2), 13–23.

Mickan, P. (2003). What is Your Score? An Investigation into language descriptors from rating written performance. *IELTS Research Reports*, *5*(3), 128–155.

Milanovic, M. (2002). *Common European framework of reference for languages: Learning, teaching, assessment: Language examining and test development*. Strasbourg: Council of Europe Language Policy Division.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic, & N. Saville (Eds.), *Language testing* 3 *–Performance, testing, cognition and assessment,* (pp. 92–114). Cambridge: Cambridge University Press.

Mislevy, R. J., & Riconscente, M. M. (2005). Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report No. 9). Menlo Park, CA: SRI and University of Maryland. Retrieved on 21 June 2020, from http://padi.sri.com/downloads/TR9_ECD.pdf.

Molnár, E. K. (2002). Az írásbeli szövegalkotás. In: Csapó B. (Ed.), *Az iskolai műveltség* (pp. 193–216). Budapest: Osiris.

Molnár, E. K. (2009). Az írásbeli szövegalkotás funkciója és hatékonysága magyar egyetemista diákok dolgozatainak szövegeiben. *Anyanyelv-pedagógia, 2*(1). Retrieved on 16 June 2020, from http://www.anyanyelv-pedagogia.hu/cikkek.php?id=138

Moore, F. M. (2007). Language in science education as a gatekeeper to learning, teaching, and professional development. *Journal of Science Teacher Education, 18*, 319–343.

Moore, Y. (2015*). Investigating valid constructs for writing tasks in EAP tests for use in Japanese university entrance examinations*. ARAGs Research Report Online: British Council. Retrieved on 16 June 2020, from: https://www.britishcouncil.org/sites/default/files/investigating_valid_constructs_for_writing_in_eap_tests_for_use_in_japanese_university_entrance_examinations.pdf

Moore, T., & Morton, J. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*, 43–66.

Morrow, K. (2004). Background to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 3–11). Oxford: Oxford University Press.

Myles, J. (2002). Second language writing and research: The writing process and error analysis in student texts. *The Electronic Journal for English as a Second Language, 6*(2), 1-12. Retrieved on 1 June, 2020 from: http://www.tesl-ej.org/wordpress/issues/volume6/ej22/ej22a1/

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education, 25*(4), 262–279.

Newton, P. E., & Baird, J. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*(2), 173–177.

Newton, P. E. (2017). *An approach to understanding validation arguments*. Coventry: Ofqual. Retrieved on 16 June 2020, from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/653070/An_approach_to_understanding_validation_arguments.pdf

North, B. (2000). *The development of a common framework scale of language proficiency.* New York, NY: Peter Lang.

North, B. (2004). Relating assessments, examinations, and courses to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework*, (pp. 77–90). Oxford: Oxford University Press.

North, B. (2014). *The CEFR in practice*. Cambridge, UK: Cambridge University Press.

Norton, L. S. (1990). Essay-writing: What really counts? *Higher Education, 20*, 411–442.

O'Sullivan, B. (2012). The assessment development process. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 47–58). New York: Cambridge University Press.

O'Sullivan, B. (2018, May). *Localisation: The ultimate goal in language testing.* Paper presented at the 15th EALTA Conference, TestDaf Institute, Bochum, Germany.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis general technical manual*. London: British Council.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33–56.

Ostler, S. (1987). English in parallels: A comparison of English and Arabic prose. In U. Connor, & R. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (pp. 169–185). Reading, MA: Addison, Wesley.

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), The *Routledge handbook of applied linguistics* (pp. 259-273). Abingdon: Routledge.

Paltridge, B., & Phakiti, A. (2015). *Research methods in applied linguistics: A practical resource.* London, UK: Bloomsbury Publishing.

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the common European framework of reference*. Frankfurt am Main, Germany: Peter Lang.

Paulsen, C. A., Levine, R. (1999, April). *The applicability of the cognitive laboratory method to the development of achievement test items*. Paper presented in Research in the Development of Tests and Test Items, at the annual meeting of the American Educational Research Association, Montreal. Retrieved on 16 June 2020, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.511.4341&rep=rep1&type=pdf

Pearson Academic (2017). *Test format*. Retrieved on 16 June 2020, from https://pearsonpte.com/the-test/format/

Pearson Academic (2019). *Score guide for test takers: Version 12 – October 2019*. Retieved from: https://pearsonpte.com/wp-content/uploads/2019/10/Score-Guide-for-test-takers-V12-20191030.pdf

Perie, M., & Huff, K. (2016). Determining content and cognitive demand. In: S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 119–143). New York and London: Routledge.

Perlman, M. (2013). Finalizing the test blueprint. In: C. A. Chapelle, M. K. Enright, & J. M. Jamieson. (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). NY, New York: Routledge.

Piccardo, E. (2018). Plurilingualism: Vision, conceptualization, and practices. In: P. P. Trifonas, & T. Aravossitas (Eds.), *Handbook of research and practice in heritage language education* (pp.1–19), Cham: Springer International Publishing.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111–129.

Plakans, L. (2012). Writing integrated items. In: G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 249–261). New York: Routledge.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural modelling equation approach* (Studies in Language Testing). Cambridge: Cambridge University Press.

Read, J. (2015). *Assessing English proficiency*. Palgrave: Macmillan.

Ringwald, C. (2018, May). *Are you ready for the transition to higher education? The preparatory power of the German Abitur for English-medium bachelor degree programs*. Paper presented at the 15th EALTA Conference, SIG Assessment of Writing/Assessment for Academic Purposes, TestDaf Institute, Bochum, Germany.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Vol. 2. Reading, writing, and language learning* (pp. 142–175). Cambridge: Cambridge University Press.

Scollon, R. (2001). Action and text: Toward an integrated understanding of the place of text in social (inter)action. In R. Wodak, & M. Meyer (Eds.), *Methods in critical discourse analysis* (pp. 139-183). London: Sage.

Shaw, S. D., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication*, Special Issue 3, 1–44.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*, Studies in Language Testing (Vol. 26). Cambridge: UCLES/Cambridge University Press.

Shi, L. (2004). Textual borrowing in second language writing. *Written Communication*, *21*(2), 171–200.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: NFLC Occasional Papers.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.

Smagorinsky, P (1994). Think-aloud protocol analysis: Beyond the black box. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 3–19). Thousand Oaks: Sage.

Spack, R. (1988). Initiating ESL students into the academic discourse community: How far should we go? *TESOL Quarterly 22*(1), 29–51.

Sperling, M. (1996). Revisiting the writing-speaking connection: Challenges for research on writing and writing instruction. *Review of Educational Research, 66*, 53–86.

Spolsky, B., Inbar-Lourie, O., & Tannenbaum, M. (Eds.) (2014). *Challenges for language education and policy. Making space for people.* New York, NY*:* Routledge.

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*, 187–201.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.) *Sociocultural theory and second language learning* (pp. 97–114). Oxford: Oxford University Press.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Taylor, L. (2002). Assessing learner's English: but whose/which English(es)? *Research Notes 10*, 18–20.

Taylor, L. (2004). Second language writing assessment: Cambridge ESOL's ongoing research agenda. *Research Notes*, *16*(1), 2.

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154–155.

Taylor, L. (2010). Introduction. *IELTS Research Report Volume 11*, 7–20.

The Hungarian National Curriculum (2012). Retrieved on 16 June 2020, from: http://www.magyarkozlony.hu/pdf/13006

TOEFL iBT (2018). *Test content.* Retrieved on 16 June 2020, from https://www.ets.org/toefl/ibt/about/content/

Trace, J., Meier, V., & Janssen, G. (2016). “I can see that”: Developing shared rubric category interpretations through score negotiation. *Assessing Writing, 30,* 32–43. https://doi.org/10.1016/j.asw.2016.08.001

Tsushima, R. (2015). Methodological diversity in language assessment research: The role of mixed methods in classroom-based language sssessment studies. *International Journal of Qualitative Methods, 14*(2), 104–121.

Turner, C. E. (2013). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1403–1417). Hoboken, NJ: Wiley-Blackwell.

University of Bedfordshire (2019). *Book chapters & journal articles*. Retrieved on 16 June 2020, from: https://www.beds.ac.uk/crella/sociocognitive/bookchapters

University of Cambridge Language Examination Services (2016). *Assessing writing performance: Level C1*. Retrieved on 16 June 2020, from: https://www.cambridgeenglish.org/images/cambridge-english-assessing-writing-performance-at-level-c1.pdf

Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal 64*(3), 314–320.

Van Moere, Alistair. (2014). Raters and ratings. In: A. J. Kunnan (Ed.), *The companion to language assessment, 3*, (pp. 1358–1374). Chichester: Wiley.

Van Moere, A., & Downey, R. (2017). Technology and artificial intelligence in language assessment. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). Boston: De Gruyter Mouton.

Varner, L. K., Roscoe, R., & McNamara, D. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research, 5(*1), 35–59. https://doi.org/10.17239/jowr-2013.05.01.2

Verhelst, N. D., Glas, C. A., & Verstralen, H. H. F. M. (1995). *One-parameter logistic model*. Arnhem: CITO.

Vígh, T. (2005). A kommunikatív tesztelmélet alapjai. *Magyar Pedagógia, 105*(4), 381–407.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wagner, E. (2010). Survey research. In B. Paltridge, & A. Phatiki (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 22–38). London: Continuum.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.

Wall, D. and Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe. Phase 2: Coping with Change*. TOEFL iBT Report iBT-05. Princeton, NJ: Educational Testing Service. Retrieved on 16 June, 2020 from: http://www.ets.org/Media/Research/pdf/RR-08-37.pdf

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 253–287.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(9), 27–55.

Weir, C. J. (1988). Construct validity. In A. Hughes, D. Porter, & C. J. Weir (Eds.), *ELT Validation Project: Proceeding of a Conference Held to Consider the ELTS Validation Project Report*. The British Council and the University of Cambridge Local Examination Syndicate.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Weir, C. J. (2005a). *Language testing and validation: an evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the common European framework for developing comparable examinations and tests. *Language Testing*, *22*, 281–300.

Weir, C. J., Vidakovic, I., & Galaczi, D. E. (2013). *Measured constructs: A history of Cambridge English examinations, 1913-2012.* Studies in Language Testing Vol. 37. Cambridge: UCLES/Cambridge University Press.

Weninger, C., & Khan, K. H.-Y. (2013). (Critical) language awareness in business communication. *English for Specific Purposes, 32*, 59–71.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400–409.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305–319. https://doi.org/10.1177/026553229301000306

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83–106.

Woodall, B. R. (2002). Language-switching: Using the first language while writing in a second language. *Journal of Second Language Writing, 11*(1), 7–28.

Yin, J. (2011). Fundamental concerns in high-stakes language testing: The case of the college English test. *Pan-Pacific Association of Applied Linguistics, 15*(2), 71–83.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research & Evaluation, 20*(5), 1–20.

Zucker, S., Sassman, C., & Case, B. J. (2004) *Cognitive labs: Technical report*. Retrieved on 16 June 2020, from: http://images.pearsonassessments.com/images/tmrs/tmrs_rg/CognitiveLabs.pdf

Appendices

Appendix 1 Preliminary investigation of the construct: semi-structured interview questions

Semi-structured interview questions - students

Age

Institution

Program

Subject

Year of study

Language of study

Level of English (self-assessment)

Level of English (exam certificate)

How did you arrange your studies abroad (i.e. in Hungary

How do you communicate in English in connection with your studies?

When you communicate in writing how do you do that? Why?

What do you communicate about in writing?

Who do you write to?

Can you recall why you initiate communication?

Give examples of written communication you took part in in the past 2 months.

Do you think your level of English is sufficient for written communication in a university context?

Do you ever get feedback from anyone on your writing? If so, what?

Do you think you were prepared for formal written communication before your university studies?

If you hold a language exam certificate, do you think the kind of writing you did for the exam is related to this kind of writing?

Semi-structured interview questions – University staff

Institution

Faculty

Department

Level of English (self-assessment)

Level of English (exam certificate)

How do you communicate in English with international students?

When you communicate in writing how do you do that? Why?

What do you communicate about in writing?

Who communicate with you in English? Why?

Can you recall why you initiate communication in English?

Give examples of written communication you took part in in the past 2 months.

Do you think university students' level of English is sufficient for written communication in a university context?

Do you ever provide feedback from anyone on your writing? If so, what?

Do you think students were prepared for formal written communication before your university studies?

Appendix 2 Specification of the Writing Tasks for the C1 Level Euroexam Academic English Test: Preliminary version for expert judgement

Writing

This assessment consists of two tasks. The assessment criteria are: task achievement, appropriacy, coherence, cohesion, grammatical range and accuracy and lexical range and accuracy.

	Duration: 60 minutes	
Task	**Component (Task name and task focus)**	**Number of questions**
1	**Transactional writing**	1
2	**Discursive writing**	1

		Skill focus	Task description	Response format
Writing 60 minutes	**Task 1**	Transactional writing. Respond to input text and produce a formal response for an intended recipient	Candidate creates a formal email of 200 words based on verbal information.	Formal writing with a clear sense of purpose, audience and format - formal email
	Task 2	Discursive writing	Candidate writes a piece of extended text of 200-250 words for general, distant audience – choice of topic given	Neutral or informal writing focusing on personal point of view e.g. argument, opinion, discussion etc

C1

CEFR descriptors

- Can express him/herself in clear, well-structured text, expressing points of view at some length.
- Can write about complex subjects in a letter, an essay or a report, underlining what he/she considers to be the salient issues.
- Can select style appropriate to the reader in mind.

Writing Task 1

Task focus	– Respond to input text and produce a formal response for an intended recipient.
Task type	– Read input text and respond appropriately formally to an intended recipient, e.g. teacher, administrative staff at a university, library staff. – Candidate to write a formal email of ca. 200 words based on two semi-authentic input texts given.
Question format	– Standard rubric with task specific additions – Task specific instructions with two specified functions – One semi-authentic input text, written or diagrammatic (leaflets, notes, letters, maps, timetables, subject reports, library cards, etc.) which give candidate clear pointers as to content of response – One text with four content points presented as the candidate's own notes – Appropriate answer sheet
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ NB It is important that it is what is elicited that is C1 – the input texts can be below C1 – Candidate must be given intended recipient. – Candidate to be clearly asked to produce a formal email. – Candidate to be asked to define, describe, elaborate, illustrate, compare and contrast, classify, cause and effect, problem and solution, justify, hypothesise, summarise report, complain suggest, give and ask for information, express stance, opinion, argument, justify a request, explain a situation to elicit the appropriate level of language and formal style. – Candidate to be asked to use two functions in the rubric. – All input texts to appear authentic. NB it must not be possible to copy the content of these as part of the candidate's answer. – The topic must be accessible to a wide range of learners.
Standard rubric	– See template provided as a separate document. Slight variation between versions.
Length	– Maximum overall word count 100 words (instructions and input texts combined)
Time	– 30 minutes recommended for task
Artwork	– The input texts should appear as authentic as possible i.e. it should follow the layout features of authentic material the candidate encounters in real life.
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 1
Checklist	– Checklist for Writing 1 to be completed and submitted with item

Writing Task 2

Task focus	– Respond to input text to produce a personal response for an intended public audience.
Task type	– Read input text and respond appropriately. Candidates choose one task from a selection of three which have different topics and require different functions and genres – Genres should be three of the following: essay (descriptive, discursive, argument), review, report, article
Question format	– Three sets of instructions which make context and response requirements clear – Appropriate answer sheet
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ – The context, function and output text type should be explicitly stated – The context, function and output text type should be explicitly stated – It should be clear that a personal response is required – It should be feasible to complete the task using ca. 250 words
Standard rubric	– See template provided as a separate document. No variation between versions.
Number of items	– 3 options
Length	– Maximum 60 words per option.
Time	– 30 minutes recommended for task
Artwork	– None
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 2
Checklist	– Checklist for Writing 2 to be completed and submitted with item

Appendix 3 Expert judgement: Questionnaire

Euroexam Academic – Level C1

- Review Template -

Based on your subject knowledge and experience, to what extent …

1. *can a given test score be interpreted as an indicator of the construct* Euroexam Academic *wishes to measure?*

- (●) inadequate construct representation
- ( ) limited construct representation
- ( ) adequate construct representation

2. *do the* Euroexam Academic *test tasks reflect the characteristics of the target language use tasks?*

- (●) inadequate domain coverage
- ( ) limited domain coverage
- ( ) adequate domain coverage

3. *do the* Euroexam Academic *test tasks provide an adequate coverage of the content of the target language?*

- (●) inadequate content coverage
- ( ) limited content coverage
- ( ) adequate content coverage

4. *do the* Euroexam Academic *test task characteristics correspond to the features of the target language use tasks?*

- (●) inadequate correspondence
- ( ) limited correspondence
- ( ) adequate correspondence

5. *do the contextual features of the* Euroexam Academic *test tasks correspond to the characteristics of the target language situation?*

- (●) inadequate situational authenticity
- ( ) limited situational authenticity
- ( ) adequate situational authenticity

6. *does the test taker need to rely on their individual characteristics in accomplishing a test task?*

- (●) inadequate interactiveness
- ( ) limited interactiveness
- ( ) adequate interactiveness

Appendix 4 Euroexam Academic English Test: List of Topics

Topics for Euroexam Academic

PERSONAL IDENTIFICATION

- title
- name
- marital status
- age
- gender
- occupation
- nationality
- first language, second language
- character
- opinion
- image

HOUSE & HOME and LOCAL ENVIRONMENT

- interior design
- local & regional services/amenities
- regional geographical features
- natural environment
- region-specific phenomena
- student accommodation
- halls of residence
- house sharing
- letting agencies
- estate agents

DAILY LIFE - WORK REALTED

- income, salary variations
- prospects
- private pursuits
- stress
- money management
- part time and full time work
- the world of work and technological development
- empowerment of women
- children at work

FREE TIME, ENTERTAINMENT

- TV, radio, cinema, theatre
- computer, internet
- intellectual/artistic pursuits
- sports
- press
- music
- photography
- reading habits, letter-writing, diaries etc
- exhibitions, museums
- leisure/work ratio
- university societies/clubs

TRAVEL

- traffic & traffic control
- travel for business and holiday purposes
- infrastructural development
- pollution and environmental issues
- 'green' travel
- entering and leaving a country
- common currency eg. the euro
- migration

SOCIETY/RELATIONS WITH OTHER PEOPLE

- family relationships
- friendship
- correspondence
- manners
- social conventions
- social life
- government and politics
- crime and justice
- war and peace
- anti-social behaviour
- generation gaps
- individual rights
- freedom of speech
- media censorship
- social responsibilities
- equal opportunities
- human rights
- citizenship
- population explosion
- government services
- crime
- prisons and rehabilitation
- poverty
- society and family
- responsibilities of parents
- role models
- materialism and consumerism
- different cultures
- Internet censorship
- social networking

HEALTH AND BODY CARE

- personal hygiene
- health and illness
- medical services
- insurance issues
- obesity
- prevention or cure

EDUCATION

- schooling
- subjects
- qualifications and examinations
- education systems
- teaching and learning
- education and technology
- education and government funds
- campuses

- university/college building facilities
- teaching/learning methods
- library systems
- online courses
- life-long learning
- career planning
- internship/work experience

SHOPPING

- shopping facilities
- foodstuffs
- household articles
- prices
- ethical shopping
- retail therapy

FOOD AND DRINK

- eating habits
- sourcing food locally
- fast food
- organic food
- year-round availability
- diets
- food fashions

SERVICES

- communications
- financial services
- emergency services
- leisure facilities
- IT in the community
- diplomatic services
- employment agencies
- business management
- business and technology

PLACES & LOCATION

- satellite navigation systems
- World Heritage sites
- roads and motorways, airports
- protecting open spaces
- how geography affects people

LANGUAGE

- foreign language ability
- accents and dialects
- preserving minority languages
- bilingualism
- multilingualism
- universal languages eg. Esperanto
- body language
- translation/interpretation
- English as a lingua franca (ELF)

WEATHER

- climate and weather
- weather forecasting
- climate change

- extreme weather
- weather and mood

NUMBERS AND TRENDS

- statistics
- processes
- importance of maths in everyday life
- interpretation of graphs
- describing trends

THE ENVIRONMENT

- recycling
- pollution
- global warming
- endangered species
- future of the planet
- animal protection
- sustainability

ARTS, SCIENCES AND SOCIAL SCIENCES

- modern art, theatre, architecture
- classical art, theatre, architecture
- literature
- popular culture
- censorship of arts and artists
- arts and academic studies
- scientific development
- space exploration
- power of the computer
- important inventions
- genetic modification
- ethics
- animal testing

Appendix 5 Specification of the Writing Tasks for the C1 Level Euroexam Academic English Test: Updated after Stage 1

Writing

This assessment consists of two tasks. The assessment criteria are: task achievement, appropriacy, coherence, cohesion, grammatical range and accuracy and lexical range and accuracy.

	Duration: 60 minutes	
Task	**Component (Task name and task focus)**	**Number of questions**
1	**Transactional writing**	1
2	**Discursive writing**	1

		Skill focus	Task description	Response format
Writing 60 minutes	**Task 1**	Transactional writing. Respond to input text and produce a formal response for an intended recipient	Candidate creates a formal letter or email of 200 words based on verbal information.	Formal writing with a clear sense of purpose, audience and format - formal email
	Task 2	Discursive writing	Candidate writes a piece of extended text of 200-250 words for general, distant audience – choice of topic given	Neutral or informal writing focusing on personal point of view e.g. argument, opinion, discussion etc

C1

CEFR descriptors

- Can express him/herself in clear, well-structured text, expressing points of view at some length.
- Can write about complex subjects in a letter, an essay or a report, underlining what he/she considers to be the salient issues.
- Can select style appropriate to the reader in mind.

Writing Task 1

Task focus	– Respond to input text and produce a formal response for an intended recipient.
Task type	– Read input text and respond appropriately formally to an intended recipient, e.g. teacher, administrative staff at a university, library staff. – Candidate to write a formal email of ca. 200 words based on two semi-authentic input texts given.
Question format	– Standard rubric with task specific additions – Task specific instructions with two specified functions – One semi-authentic input text, written or diagrammatic (leaflets, notes, letters, maps, timetables, subject reports, library cards, etc.) which give candidate clear pointers as to content of response – One text with four content points presented as the candidate's own notes – Appropriate answer sheet
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ NB It is important that it is what is elicited that is C1 – the input texts can be below C1 – Candidate must be given intended recipient. – Candidate to be clearly asked to produce a formal email. – Candidate to be asked to define, describe, elaborate, illustrate, compare and contrast, classify, cause and effect, problem and solution, justify, hypothesise, summarise report, complain suggest, give and ask for information, express stance, opinion, argument, justify a request, explain a situation to elicit the appropriate level of language and formal style. – Candidate to be asked to use two functions in the rubric. – All input texts to appear authentic. NB it must not be possible to copy the content of these as part of the candidate's answer. – The topic must be accessible to a wide range of learners.
Standard rubric	– See template provided as a separate document. Slight variation between versions.
Length	– Maximum overall word count 100 words (instructions and input texts combined)
Time	– 30 minutes recommended for task
Artwork	– The input texts should appear as authentic as possible i.e. it should follow the layout features of authentic material the candidate encounters in real life.
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 1
Checklist	– Checklist for Writing 1 to be completed and submitted with item

Writing Task 2

Task focus	– Respond to input text to produce a personal response for an intended public audience.
Task type	– Read input text and respond appropriately. Candidates choose one task from a selection of three – The genre is discussion essay in three distinct fields of study.
Question format	– Three sets of instructions which make context and response requirements clear – Appropriate answer sheet
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ – The context, function and output text type should be explicitly stated – The fields of study for the three questions should always be the following: o humanities/social science o science o business/economy – It should be clear that a personal response is required – It should be feasible to complete the task using 200-250 words
Standard rubric	– See template provided as a separate document. No variation between versions.
Number of items	– 3 options
Length	– Maximum 60 words per option.
Time	– 30 minutes recommended for task
Artwork	– None
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 2
Checklist	– Checklist for Writing 2 to be completed and submitted with item

Appendix 6 Accredited C1 level Euroexam writing assessment scale

Euro C1: Writing Mark Scheme

	Task Achievement	Appropriacy	Coherence	Cohesion	Grammatical Range and Accuracy	Lexical Range & Accuracy
5	**Task achieved at a high level Intention** **Intention**: Entirely clear **Instructions**: Completely followed **Effect**: A positive effect on the target reader **Outcome**: Sure to achieve a successful outcome **Content**: All relevant details included. Some original ideas or presentation	**Style & Format**: Appropriate to genre, no irrelevant information **Register**: Good awareness of register and formality level appropriate to genre	**Structure**: Ideas sequenced logically and accurately **Purpose**: Clear **Information**: Well organised into a coherent text	**Grammatical Structures**: A wide range of cohesive devices used naturally, efficiently and appropriately to link words, clauses, sentences and paragraphs **Reference**: Skilled use	**Grammatical Structures**: Complex **Spelling**: Very good **Word order**: Correct **Punctuation**: Used properly throughout **Errors**: Very few, none of them impedes meaning, message	Wide range of lexis to complete the task, some original lexical solutions Lexis used appropriately with isolated misuse
4						
3	**Task achieved, some gaps** **Intention**: Clear in most areas **Instructions**: All important ones followed **Effect**: A generally positive effect on the reader. **Outcome**: Likely to achieve a successful outcome **Content**: Many relevant details included	**Style & Format**: Usually appropriate to genre with little or no irrelevant information **Register**: Limited exponents but awareness of register is shown	**Structure**: Some confusion in logical and accurate sequencing **Purpose**: Mostly clear **Information**: Adequately organised into a mostly coherent text	**Grammatical Structures**: Adequate amount of devices used to link words, clauses, sentences mostly appropriately **Reference**: Limited and inaccurate use	**Grammatical structures**: Adequately complex structures with rare mistakes that do not impede comprehension **Spelling**: Some mistakes that do not impede comprehension **Word order**: Mostly correct. **Punctuation**: Mostly effective **Errors**: Some, but do not significantly impede meaning.	Sufficient range of lexis to complete the task Lexis used mostly appropriately with some occasional misuse
2						
1	**Task unachieved** **Intention**: Very unclear. **Instructions**: Many not followed **Effect**: Negative **Outcome**: Will not achieve a successful outcome **Content**: Omission, irrelevance.	**Style & Format**: Inappropriate to genre, or minimal evidence **Register**: Minimal	**Structure**: Muddled **Purpose**: Unclear **Information**: Very confused	**Grammatical Structures**: Minimal **Reference**: Simple / none	**Grammatical Structures**: Very simple with frequent and serious mistakes **Spelling**: Very poor **Word order**: Often wrong **Punctuation**: Often wrong	Poor range of lexis to complete the task Lexis used inappropriately in most cases
0	Task unattempted / partially attempted Not enough language to make an assessment, or under 20 words.	Not enough language to make an assessment, or under 20 words	No meaning or the meaning conveyed is irrelevant, or under 20 words	No effective use of cohesive devices and reference, or under 20 words	Little or no evidence of grammatical knowledge of simple structures, or under 20 words.	No relevant lexis organized into sentences, or under 20 words.

euroexam international

Appendix 7 Stage 2 Domain modelling and trialling: semi-structured interview questions

(The interviews were conducted in Hungarian; the questions in the Appendix appear in my translation.)

Age

Institution

Program

Subject

Year of study

Language of study

Level of English (self-assessment)

Level of English (exam certificate)

How did you start completing the Writing Paper?

How did you feel when completing the Writing Paper?

How did you plan your writing?

How did you decide on the structure of the texts? Formal email – essay.

Do you think your vocabulary was sufficient to complete the tasks? Formal email – essay.

How would you compare the two tasks?

Did you use a dictionary? Why/Why not?

How long did you spend on each task?

Do you think the allocated time is sufficient?

Appendix 8 Specification and sample tasks of the Writing Tasks for the C1 Level Euroexam Academic English Test: Updated after Stage 2

Writing

This assessment consists of two tasks. The assessment criteria are: task achievement, appropriacy, coherence, cohesion, grammatical range and accuracy and lexical range and accuracy.

	Duration: 60 minutes	
Task	**Component (Task name and task focus)**	**Number of questions**
1	**Transactional writing**	1
2	**Discursive writing**	1

		Skill focus	Task description	Response format
Writing 60 minutes	**Task 1**	Transactional writing. Respond to input text and produce a formal response for an intended recipient	Candidate creates a formal email of 200 words based on verbal information.	Formal writing with a clear sense of purpose, audience and format - formal email
	Task 2	Discursive writing	Candidate writes a piece of extended text of 200-250 words for general, distant audience – choice of topic given	Neutral or informal writing focusing on personal point of view e.g. argument, opinion, discussion etc

C1

CEFR descriptors

- Can express him/herself in clear, well-structured text, expressing points of view at some length.
- Can write about complex subjects in a letter, an essay or a report, underlining what he/she considers to be the salient issues.
- Can select style appropriate to the reader in mind.

Writing Task 1

Task focus	– Respond to input text and produce a formal response for an intended recipient.
Task type	– Read input text and respond appropriately formally to an intended recipient, e.g. teacher, administrative staff at a university, library staff. – Candidate to write a formal email of ca. 200 words based on two semi-authentic input texts given.
Question format	– Standard rubric with task specific additions – Task specific instructions with two specified functions – One semi-authentic input text, written or diagrammatic (leaflets, notes, letters, maps, timetables, subject reports, library cards, etc.) which give candidate clear pointers as to content of response – One text with four content points presented as the candidate's own notes – Appropriate answer sheet with email layout
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ NB It is important that it is what is elicited that is C1 – the input texts can be below C1 – Candidate must be given intended recipient. – The name of intended recipient must be provided. – Candidate to be clearly asked to produce a formal email. – Candidate to be asked to define, describe, elaborate, illustrate, compare and contrast, classify, cause and effect, problem and solution, justify, hypothesise, summarise report, complain suggest, give and ask for information, express stance, opinion, argument, justify a request, explain a situation to elicit the appropriate level of language and formal style. – Candidate to be asked to use two functions in the rubric. – All input texts to appear authentic. NB it must not be possible to copy the content of these as part of the candidate's answer. – The topic must be accessible to a wide range of learners.
Standard rubric	– See template provided as a separate document. Slight variation between versions.
Length	– Maximum overall word count 100 words (instructions and input texts combined)
Time	– 30 minutes recommended for task
Artwork	– The input texts should appear as authentic as possible i.e. it should follow the layout features of authentic material the candidate encounters in real life. – Email template to be provided on answer sheet.
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 1
Checklist	– Checklist for Writing 1 to be completed and submitted with item

Writing Task I example

You are not satisfied with the mark and evaluation you have been given for your college course. Write an email to your teacher explain what happened and justify your request. Use the information below.

Write 200 words.

Student evaluation form

Course title: Theory of Knowledge

Teacher: Professor Peter Johannsen

Time: Monday 14:00-16:00

Attendance: Poor

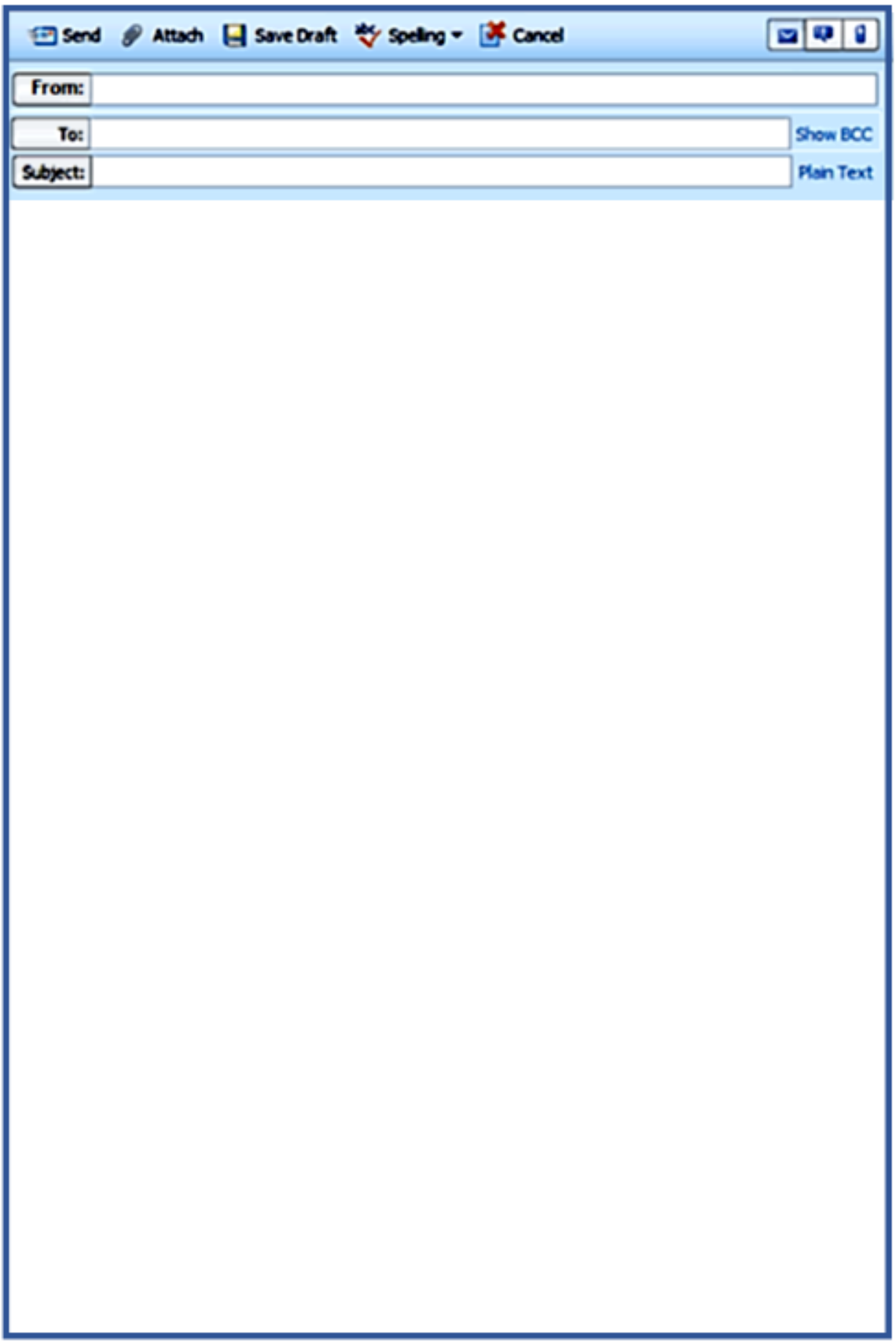
Midterm test (underlined): Fail – Pass – Average – Good - Excellent

Essay: Not handed in

Final mark (underlined): Fail – Pass – Average – Good - Excellent

My notes

- *reason for poor class attendance*
- *individual preparation for midterm - good results*
- *issues with essay*
- *new deadline for essay?*


Send
Attach
Save Draft
Spelling
Cancel
From:
To:
Show BCC
Subject:
Plain Text

Writing Task 2

Task focus	– Respond to input text to produce a personal response for an intended public audience.
Task type	– Read input text and respond appropriately. Candidates choose one task from a selection of three – The genre is discussion essay in three distinct fields of study.
Question format	– Three sets of instructions which make context and response requirements clear – Appropriate answer sheet
Question requirements	– The grammar and vocabulary used and elicited should be suitable for this level – see grammar grids below and English Vocabulary Profile vocabulary.englishprofile.org/ – The context, function and output text type should be explicitly stated – The fields of study for the three questions should always be the following: o humanities/social science o science o business/economy – It should be clear that a personal response is required – It should be feasible to complete the task using 200-250 words
Standard rubric	– See template provided as a separate document. No variation between versions.
Number of items	– 3 options
Length	– Maximum 60 words per option.
Time	– 30 minutes recommended for task
Artwork	– None
Dictionary allowed	– Yes
Mark scheme	– Mark scheme for Academic writing Task 2
Checklist	– Checklist for Writing 2 to be completed and submitted with item

Writing Task 2 example

Task Two: Discursive Writing *(30 minutes)*

- Choose only **ONE** of the following questions – 1, 2 OR 3.
- Write 200-250 words.
- **DO NOT** answer more than one question.
- Write your answer to this question on the Answer Sheet -Task Two

1. To what extent do you agree with the statement: "Participation in a student government greatly benefits your future career." Write an *essay*. Explain your points for and against and provide a conclusion at the end. Make sure you state your arguments in a logical way.

2. To what extent do you agree with the statement: "It is impossible to improve people's standard of living without using non-renewable energy resources." Write an *essay*. Explain your points for and against and provide a conclusion at the end. Make sure you state your arguments in a logical way.

3. To what extent do you agree with the statement: "The easy money effect of credit cards stimulates overspending." Write an *essay*. Explain your points for and against and provide a conclusion at the end. Make sure you state your arguments in a logical way.

Appendix 9 Test Taker Questionnaire for the Euroexam Academic Pretest

Euroexam International Academic English
ELŐTESZTELÉS
Vizsgázói adatlap és véleményezőlap

	SZEMÉLYES ADATOK			
Neve				
Születési ideje (éééé/hh/nn)				
Neme (F/N)				
Iskolája neve				
Mikor kezdett angolul tanulni? (évszám)				
Milyen szintű nyelvvizsgája van angolból? Karikázza be a szintet, és adja meg a vizsgarendszert (pl. Euro, Origó, stb.)	**B1** **vizsgarendszer:**	**B2** **vizsgarendszer:**	**C1** **vizsgarendszer:**	**egyéb** **szint:** **vizsgarendszer:**
	A VIZSGÁRA VONATKOZÓ VÉLEMÉNYEK			
A vizsga mely részét találta legnehezebbnek? Miért?				
A vizsga mely részét találta legkönnyebbnek? Miért?				
Mennyi időre volt szüksége a feladatok megoldásához?	**Reading** ..... **perc**	**Listening** ..... **perc**	**Writing** ..... **perc**	
Érzése szerint elér 40%-ot vagy annál többet az egyes feladatokon?				
Érzése szerint elér 60%-ot vagy annál többet az egyes feladatokon?				

Kérjük, értékelje a tesztet (1= nagyon rossz 5= kiváló) az alábbiak alapján:

	1	2	3	4	5
Megjelenés és kivitelezés					
Utasítások					
Feladattípusok					
Tartalom					

Appendix 10 Test taker performance sample: fail

WRITING – **ANSWER SHEET – TASK TWO**

3. "Getting a University education is no longer a guarantee of success."

Many people attend universities in different fields, but there are others, for whom university is not their cup of tea. However, what if your long-expected diploma does not count so much anymore?

First of all, reaching university seems to be unbelievable, either because they do not even want to study more, rather to have a job; or they unfortunately don't score enough to get in. No matter which group we are talking about, these people have only got a slight chance to work as managers of their own company. But then, what about those who get into the university?

While in the past high-school education meant a guarantee to find a job that might provides a good salary, these days it is out of fashion. Holding a Ma/MSC degree in your hand while looking for a good-enough job matters. On the other hand, people frequently can't make head or tail of their Ba/BSc diploma to get a job.

Opportunities tend to meet those who have better relationships, so giving a job or a place to an acquaintance is not an uncommon phenomenon.

Moreover, there are some world-known highschools, where the graduation equals a high-class job with good value. These institutions provide their students plenty of practice beside the theoretical lessons, in order to "make them worth in the market".

All things considered, it is worth attending a highschool to obtain deeper knowledge of a particular field. The more skills and knowledge you have, the more chance you've got to get your dream job.

Appendix 11 Test taker performance sample: pass

WRITING – **ANSWER SHEET – TASK TWO**

Essay about "Getting a university degree is no longer a guarantee of success"

First of all ~~In my point of view~~, nowadays getting a university degree is not a must, however it could mean a lot to your future employer. At present of course, ~~it is an~~ ~~available~~ you can choose whether you want to go to a university, or not.

However, having a university degree ~~could~~ can increase the amount of money you would get ~~and the~~ at the ~~end~~ of the month. ~~Although~~ But, what do we consider as a success? It could be pleasant to do anything without a degree as well. In addition a person could work at from a younger age. The young might learn from the experiences they got from their previous workplace. On the other hand university students usually form into different groups after they leave our educational system. I mean, they more likely to be with each other.

Additionally, these connections made ~~in~~ ~~unwant~~ while ~~after~~ attending a university, can help to ~~toda~~ achive life-goals much more easily.
For example, ~~if~~ if your child is sick or ill, you can get help from a friend faster.
~~As far as I can see~~
As far as I can see getting a degree is the biggest opportunity you can have.
Do not miss it.

Appendix 12 Test taker performance sample: pass with distiction

WRITING – **ANSWER SHEET – TASK TWO**

3, Getting a university education is no longer a guarantee of success

Unaffordable tuition fees, great opportunities and new friends. I think these are one of the few things which instantly come to one's mind when someone mentions university. A vast majority of my friends will certainly go to universities, but when I asked why they simply said: "It is a must!" So, is it really? I think just simply getting an university education is not quite a guarantee of success.

Each and every year thousands and thousands of people apply their studies, but only a chosen few, with the highest points achieved - are eligible for state funded courses. Tuition fees have been skyrocketing lately, but not just on a national scale. (~~I think it is safe to claims that fees have grown all around Europe.~~) Getting a diploma in some cases or in certain fields is not even needed, for example you can be the best journalist or locksmith in the country and you spend your valuable years not by studying, but by mastering all aspects of your field.

You can even become a billionaire. Who has not heard of Bill Gates and how he built himself and Microsoft up after dropping out of college. (~~Now, I know he is one of a kind, and by no means is the route to becoming a billionaire as easy as dropping out of college.~~)

Euroexam Level C1 - Writing - ANSWER SHEET

But yes, there are fields where a diploma will get you far. For example by studying law, economy or by becoming a doctor your route to success is paved. OECD finds that diploma holders are more likely to earn more, and less likely to lose their jobs when a depression hits the world and national economies. They also live longer and live a more self-conscious life. If you've chosen your studies well, I can't see a burden why you couldn't pay the tuition fees easily.

All in all I don't think university education is needed for success, but by choosing well you have a great chance of making it in life. After all you make ~~your~~ your own luck, not schools you attended. A diploma itself probably never was the guarantee of success, you are.

Appendix 13 Checklist: Preliminary version – 34 items

The logic of the statements is not punishment or reward but merely noticing the presence or absence of a feature.
The reference scale headings in the middle serve as a direct link to the CEFR (2018).
The concept check questions on the right serve to ease the decision-making process.
If the answer to all the questions in a cell is in the positive, allocate a 1 indicating that the target trait is present. If there is a negative, answer, allocate a 0.

Statement	CEFR reference scale	Concept check questions
1. This text is **legible**, i.e. the reader doesn't have to guess what the writer is trying to say.	orthographic control	Can you read the text without having to re-read words? Is the text legible? Can you keep your role as reader?
2. This text follows the standard **layout** of an essay.	orthographic control	Does the script follow standard paragraphing conventions? Are there at least four visible paragraphs (intro, more than one body paragraph, conclusion)?
3. This text is **clear** and concise.	overall written production	Are there signs of planning so that the reader's work is easier? Can you keep your role as reader? Can the writer employ the structure and conventions of the genre?
4. **Spelling** is consistently accurate.	orthographic control	Are there only **two or fewer** spelling mistakes?
5. **Punctuation** errors don't lead to misunderstanding.	orthographic control	Is the script free from punctuation errors?
6. This text is the required length as defined by the task.	TASK	Is the script within the acceptable bounds as defined by the task (in this case cca.250 words)?
7. This text displays **situational authenticity** and **self-disclosure**, i.e. the writer gives the necessary details for contextualisation, such as role, location, situation, domain, etc.	creative writing & correspondence	Does the writer explicitly state their role, location, situation, domain, etc.? Does the reader get detailed realistic information about the writer's context, situational underpinnings, setting, etc.?
8. The writer can explain the background to the [**problem**].	pluricultural repertoire	Can the writer reflect upon cultural values and practices? Can the writer deal with ambiguity in cross-cultural communication? Are the writer's reactions expressed constructively and culturally appropriately?
9. This text could **function** as a real life essay.	creative writing	Does the script clearly state (a) the reason [*why getting a*

		university education is no longer a guarantee of success] and (b) the details specified by the rubric?
10. The writer clearly articulates the [**problem**].	creative writing	Does the script contain the presentation of the problem as set in the task instructions? Is it following established conventions of the genre concerned in clear, well-structured, smoothly flowing text?
11. The script contains genuine ideas.	creative writing	Does the script contain well-structured and developed descriptions? Is the text imaginative? Is the text presenting unexpected content?
12. The writer can expand and support points of view.	creative writing	Are the paragraphs of approximately the same length? Does each paragraph contain subsidiary points?
13. The writer can present **multiple points of view**.	creative writing	Does the text contain reasons and relevant examples? Is the writer offering recommendations?
14. There is little or no **irrelevant information** in this text.	thematic development	Is the script limited to information that is directly related to the topic as defined by the task rubric?
15. The writer can hold the reader's attention and communicate complex ideas.	thematic development	Is there logical coherence in the text? Is there clear progression? Are the key ideas clearly expressed?
16. The content elements required by the task instructions are elaborated in appropriate **detail**.	TASK/thematic development	Are all the content elements discussed? Are these elaborated beyond being merely mentioned? Are they discussed in at least **one paragraph** each?
17. This text would make the appropriate **effect** on the intended audience.	planning & sociolinguistic appropriateness	Does the reader know what the writer's purpose is? Does the script clearly state what the writer [discusses]?
18. The writer adopts the level of **formality** adequate to the topic, agents, situation, domain, etc.	sociolinguistic appropriateness	Are sensitive topics handled with care? Is the writer tactful enough so that the reader is not offended? Is the level of formality adequate to the communicative context and its agents?
19. Each **paragraph** presents one distinct and unified idea.	coherence and cohesion	Does each paragraph focus on one idea? Is each **body paragraph** longer than a single sentence? If not, is it complex enough to realize a paragraph?

20. Each paragraph contains a topic sentence.	coherence and cohesion	Are the topic sentences relevant and meaningful? Do the topic sentences make it clear why the points are important?
21. The paragraphs create a **logically structured text** that is easy for the reader to follow.	coherence and cohesion	Are the paragraphs linked in meaning as the script unfolds? Is there a semantic link over and above the presentation of a list? Is there a thematic development leading the reader from the introduction through the details to the conclusion?
22. This text shows a variety of cohesive devices.	coherence and cohesion	Does the writer demonstrate the controlled use of connectors and organisational patterns?
23. This text deploys a **range** of grammatical **structures**, including the **tenses and aspects.** And **modality**.	general linguistic range	Can the writer indicate shifts in time? Is there more than one tense used? Are there any other structures (e.g. "be going to" or similar verb phrases, gerund/infinitive, non-finite verbs, conditional structures, indirect questions, reported speech, etc.)? Are grammatical aspects used consistently well?
24. The text is characterised by a broad range of language.	general linguistic range	Can the writer express him/herself clearly, without having to restrict what he/she wants to say?
25. The text demonstrates advanced vocabulary & word order and varying sentence length.	flexibility	Can the writer make a positive impact on an intended audience by effectively varying style of expression?
26. The text shows a good control of uncommon lexical items.	flexibility & sociolinguistic appropriateness	Does the text show a natural and sophisticated use of lexical features?
27. This text uses **a range of discourse functions** in a meaningful way.	turntaking	Does the script contain stock phrases to preface the writers remarks? Are these used in a meaningful way?
28. The style and tone of the text is appropriate.	vocabulary control/flexibility	Does the style and tone of the text demonstrate an appropriate control of the vocabulary of academic topics?
29. The writer can use the English **lexicon** to express the intended meaning instead of periphrases or non-existent terms.	vocabulary control	Are **all** the words existing English items? Are **all** the words used in correct meanings? Does the writer find the correct word instead of relying on periphrasis? Does the writer find the correct word instead of creating one?

		Does the text demonstrate that the writer does not have to restrict what he/she wants to say?
30. This text uses **collocations and idiomatic expressions**.	vocabulary control	Is the language of the text expressive? Is the text characterised by idiomaticity? Are the idioms used meaningfully? Are there only minor slips?
31. This text demonstrates that the writer can use **complex** sentence structures.	grammatical accuracy	Can the writer use complex sentence forms to express his/her ideas?
32. Grammatical or linguistic **errors** in this text are difficult to spot.	grammatical accuracy	Is the script free from grammatical errors that could lead to misunderstanding? Is the script free from other linguistic errors that could lead to misunderstanding?
33. This text makes use of linguistic modality.	propositional precision/grammatical accuracy	Can the writer make effective use of linguistic modality to signal the strength of a claim, an argument or a position? Are certainty/uncertainty, belief/doubt, likelihood expressed in an effectice way? Are all modal verby used consistently well? Are all the modal verbs used with the bare infinitive?
34. There is no **L1 or L3 interference** that makes reading difficult.	grammatical accuracy	Does the language of the script follow the rules of English at C1?

Appendix 14 Checklist: Pilot 1 version – 33 items

The logic of the statements is not punishment or reward but merely noticing the presence or absence of a feature.
The reference scale headings in the middle serve as a direct link to the CEFR (2018).
The concept check questions on the right serve to ease the decision-making process.
If the answer to all the questions in a cell is in the positive, allocate a 1 indicating that the target trait is present. If there is a negative answer, allocate a 0.

Statement	CEFR reference scale	Concept check questions
1. This text is **legible**, i.e. the reader doesn't have to guess what the writer is trying to say.	orthographic control	Can you read the text without having to re-read words? Is the text legible? Can you keep your role as reader?
2. This text follows the standard **layout** of an essay.	orthographic control	Does the text follow standard paragraphing conventions? Are there at least four visible paragraphs (intro, more than one body paragraph, conclusion)?
3. This text is **clear** and **concise**.	overall written production	Are there signs of planning so that the reader's work is easier? Can you keep your role as reader? Can the writer employ the structure and conventions of the genre?
4. **Spelling** is consistently accurate.	orthographic control	Are there only **two or fewer** spelling mistakes?
5. **Punctuation** is consistently accurate.	orthographic control	Is the text free from punctuation errors?
6. This text is the required length as defined by the task.	TASK	Is the text within the acceptable bounds as defined by the task (in this case cca.250 words)?
7. This text displays **situational authenticity** and **self-disclosure**, i.e. the writer gives the necessary details for contextualisation, such as role, location, situation, domain, etc.	creative writing & correspondence	Does the writer explicitly state their role, location, situation, domain, etc.? Does the reader get detailed realistic information about the writer's context, situational underpinnings, setting, etc.?
8. The writer can explain the background to the [**problem**].	pluricultural repertoire	Can the writer reflect upon cultural values and practices? Can the writer deal with ambiguity in cross-cultural communication? Are the writer's reactions expressed constructively and culturally appropriately?
9. This text could **function** as a real life essay.	creative writing	Does the text clearly state (a) the reason [*why getting a*

		university education is no longer a guarantee of success] and (b) the details specified by the rubric?
10. The text contains genuine ideas.	creative writing	Does the text contain well-structured and developed descriptions? Is the text imaginative? Is the text presenting unexpected content?
11. The writer can present **multiple points of view**.	creative writing	Can the writer evaluate problems and proposals? Does the text contain reasons and relevant examples? Is the writer offering recommendations? Are there any *inconsistencies in thinking or controversies* highlighted?
12. There is little or no **irrelevant information** in this text.	thematic development	Is the text limited to information that is directly related to the topic as defined by the task rubric?
13. The writer can hold the reader's attention and communicate complex ideas.	thematic development	Is there logical coherence in the text? Is there clear progression? Are the key ideas clearly expressed?
14. The content elements required by the task instructions are elaborated in appropriate **detail**.	TASK/thematic development	Are all the content elements discussed? Are these elaborated beyond being merely mentioned? Are they discussed in at least **one paragraph** each?
15. This text would make the appropriate **effect** on the intended audience.	planning/sociolinguistic appropriateness	Does the reader know what the writer's purpose is? Does the text clearly state what the writer [discusses]?
16. The writer adopts the level of **formality** adequate to the topic, agents, situation, domain, etc.	sociolinguistic appropriateness	Are sensitive topics handled with care? Is the writer tactful enough so that the reader is not offended? Is the level of formality adequate to the communicative context and its agents?
17. Each **paragraph** presents one distinct and unified idea.	coherence and cohesion	Are the paragraphs of approximately the same length? Does each paragraph focus on one idea? Is each **body paragraph** longer than a single sentence? If not, is it complex enough to realize a paragraph?
18. Each paragraph contains a topic sentence.	coherence and cohesion	Are the topic sentences relevant and meaningful? Do the topic sentences make it clear why the points are important?
19. The paragraphs create a **logically structured text** that is easy for the reader to follow.	coherence and cohesion	Are the paragraphs linked in meaning as the text unfolds? Is there a semantic link over and above the presentation of a list?

		Is there a thematic development leading the reader from the introduction through the details to the conclusion?
20. This text shows a variety of cohesive devices.	coherence and cohesion	Does the writer demonstrate the controlled use of connectors and organisational patterns?
21. The text is characterised by a broad range of language**.**	general linguistic range	Do complex structures (gerund/infinitive, non-finite verbs, conditional structures, indirect questions, reported speech, adverb phrases, modals, etc.) characterise the text?
22. This texts demonstrates the use of **tenses and aspects.**	general linguistic range	Can the writer indicate shifts in time? Is there more than one tense used? Are grammatical aspects used consistently well?
23. This text uses language to formulate thoughts precisely.	general linguistic range/flexibility	Can the writer express him/herself clearly, without having to restrict what he/she wants to say?
24. This text shows full consistency in the use of proforms.	general linguistic range	Is referencing used consistently well?
25. This text demonstrates that the writer can use **complex** sentence structures.	grammatical accuracy	Can the writer use complex sentence forms to express his/her ideas?
26. Grammatical or linguistic **errors** in this text are difficult to spot.	grammatical accuracy	Is the text free from grammatical errors that could lead to misunderstanding? Is the text free from other linguistic errors that could lead to misunderstanding?
27. The text demonstrates advanced vocabulary & word order and varying sentence length.	flexibility	Can the writer make a positive impact on an intended audience by effectively varying style of expression?
28. The text shows a good control of uncommon lexical items.	flexibility/sociolinguistic appropriateness	Does the text show a natural and sophisticated use of lexical features?
29. This text uses **a range of discourse functions** in a meaningful way.	turntaking	Does the text contain stock phrases to preface the writer's remarks? Are these used in a meaningful way?
30. The style and tone of the text is appropriate.	vocabulary control/flexibility	Does the style and tone of the text demonstrate an appropriate control of the vocabulary of academic topics?
31. The writer can use the English **lexicon** to express the	vocabulary control	Are **all** the words existing English items?

intended meaning instead of periphrases or non-existent terms.		Are **all** the words used in correct meanings? Does the writer find the correct word instead of relying on periphrasis? Does the writer find the correct word instead of creating one? Does the text demonstrate that the writer does not have to restrict what he/she wants to say?
32. This text uses **collocations and idiomatic expressions**.	vocabulary control	Is the language of the text expressive? Is the text characterised by idiomaticity? Are the idioms used meaningfully? Are there only minor slips?
33. This text makes effective use of linguistic modality.	propositional precision/grammatical accuracy	Can the writer make effective use of linguistic modality to signal the strength of a claim, an argument or a position? Are certainty/uncertainty, belief/doubt, likelihood expressed in an effective way? Are all modal verbs used consistently well? Are all the modal verbs used with the bare infinitive?

Appendix 15 Checklist: Pilot 2 version – 34 items

The logic of the statements is not punishment or reward but merely noticing the presence or absence of a feature.
The reference scale headings in the middle serve as a direct link to the CEFR (2018).
The concept check questions on the right serve to ease the decision-making process.
If the answer to all the questions in a cell is in the positive, allocate a 1 indicating that the target trait is present. If there is a negative answer, allocate a 0.

Statement	CEFR reference scale	Concept check questions
1. This text is **legible**, i.e. the reader doesn't have to guess what the writer is trying to say.	orthographic control	Can you read the text without having to re-read words? Is the text legible? Can you keep your role as reader?
2. This text follows the standard **layout** of an essay.	orthographic control	Does the text follow standard paragraphing conventions? Are there at least four visible paragraphs (intro, more than one body paragraph, conclusion)?
3. This text is **clear** and **concise**.	overall written production	Does the text give a positive overall impression? Are there signs of planning so that the reader's work is easier? Can you keep your role as reader?
4. **Spelling** is consistently accurate.	orthographic control	Are there only **two or fewer** spelling mistakes?
5. **Punctuation** is consistently accurate.	orthographic control	Is the text free from punctuation errors?
6. This text is the required length as defined by the task.	TASK	Is the text within the acceptable bounds as defined by the task (in this case cca.250 words)?
7. This text displays **situational authenticity** and **self-disclosure**, i.e. the writer gives the necessary details for contextualisation, such as role, location, situation, domain, etc.	creative writing & correspondence	Does the writer explicitly state their role, location, situation, domain, etc.? Does the reader get detailed realistic information about the writer's context, situational underpinnings, setting, etc.?
8. The writer can explain the background to the [**problem**].	pluricultural repertoire	Can the writer reflect upon cultural values and practices? Can the writer deal with ambiguity in cross-cultural communication? Are the writer's reactions expressed constructively and culturally appropriately?
9. This text fulfils the task requirements.	TASK	Does the text clearly state (a) the reason [*why getting a university education is no longer a guarantee of success*] and (b)

		the details specified by the rubric?
10. The content elements required by the task instructions are elaborated in appropriate **detail**.	TASK/thematic development	Are all the content elements discussed? Are these elaborated beyond being merely mentioned? Are they discussed in at least **one paragraph** each?
11. There is little or no **irrelevant information** in this text.	TASK/thematic development	Is the text limited to information that is directly related to the topic as defined by the task rubric?
12. The text contains genuine ideas.	creative writing	Does the text contain well-structured and developed descriptions? Is the text imaginative? Is the text presenting unexpected content?
13. The writer can present **multiple points of view**.	creative writing	Can the writer evaluate problems and proposals? Does the text contain reasons and relevant examples? Is the writer offering recommendations? Are there any **inconsistencies** or **controversies** highlighted?
14. The writer can communicate complex ideas in a logical way.	thematic development	Is there logical coherence in the text? Is there clear progression throughout? Are the key ideas clearly expressed in **each** paragraph?
15. This text would make the appropriate **effect** on the intended audience.	planning/sociolinguistic appropriateness	Does the reader know what the writer's purpose is? Does the text clearly state what the writer [discusses]? Is the **writer's position** clear from the text?
16. The writer adopts the level of **formality** adequate to the topic, agents, situation, domain, etc.	sociolinguistic appropriateness	Are sensitive topics handled with care? Is the writer tactful enough so that the reader is not offended? Is the level of formality adequate to the communicative context and its agents?
17. Each **paragraph** presents one distinct and unified idea.	coherence and cohesion	Are the paragraphs of approximately the same length? Does each paragraph focus on one idea? Is each **body paragraph** longer than a single sentence? If not, is it complex enough to realize a paragraph?
18. Each paragraph contains a topic sentence.	coherence and cohesion	Are **all** the topic sentences relevant and meaningful? Do the topic sentences make it clear why the points are important?
19. The paragraphs create a **logically structured text** that is easy for the reader to follow.	coherence and cohesion	Are **all** the paragraphs linked in meaning as the text unfolds? Is there a semantic link over and above the presentation of a list?

		Is there a thematic development leading the reader from the introduction through the details to the conclusion?
20. This text shows a variety of cohesive devices.	coherence and cohesion	Does the writer demonstrate the controlled use of connectors and organisational patterns **in each paragraph**?
21. The text is characterised by complex grammatical structures.	general linguistic range	Do complex structures (gerund/infinitive, non-finite verbs, reported speech, indirect questions, adverb phrases, modals, modals in the past, etc.) **characterise** the text?
22. This texts demonstrates the use of **tenses and aspects.**	general linguistic range	Can the writer indicate shifts in time? Is there more than one tense used? Are the tenses used consistently well? Are grammatical aspects used consistently well?
23. This text uses language to formulate thoughts precisely.	general linguistic range/flexibility	Can the writer express him/herself clearly, without having to restrict what he/she wants to say? Can the writer express **finer shades** of meaning?
24. This text shows full consistency in the use of proforms.	general linguistic range	Is referencing used consistently well? Is the **noun-pronoun** agreement used consistently well?
25. This text shows full consistency in the use of grammatical agreement.		Is **subject-verb** agreement used consistently well?
26. This text demonstrates that the writer can use **complex sentence** structures.	grammatical accuracy	Can the writer use complex sentence forms to express his/her ideas? Are there **relative clauses** and **conditional clauses** in the text? Are they used **consistently** well?
27. Grammatical or linguistic **errors** in this text are difficult to spot.	grammatical accuracy	Is the text free from grammatical errors that could lead to misunderstanding? Is the text free from other linguistic errors that could lead to misunderstanding?
28. The text demonstrates the use of advanced word order and varying sentence length.	flexibility	Can the writer make a positive impact on the intended audience by effectively varying style of expression?
29. The text shows a good control of uncommon lexical items.	flexibility/sociolinguistic appropriateness	Does the text show a natural and sophisticated use of lexical features?
30. This text uses **a range of discourse functions** in a meaningful way.	turntaking	Does the text contain stock phrases (**at least 2**) to preface the writer's remarks?

		Are these used in a meaningful way?
31. The style and tone of the text is appropriate.	vocabulary control/flexibility	Does the style and tone of the text demonstrate an appropriate control of the vocabulary of academic topics?
32.The writer can use the English **lexicon** to express the intended meaning instead of periphrases or non-existent terms.	vocabulary control	Are **all** the words existing English items? Are **all** the words used in correct meanings? Does the writer find the correct word instead of relying on periphrasis? Does the writer find the correct word instead of creating one? Does the text demonstrate that the writer does not have to restrict what he/she wants to say?
33. This text is characterised by the use of **collocations and idiomatic expressions**.	vocabulary control	Is the language of the text expressive? Is the text characterised by idiomaticity? Are the idioms used meaningfully? Are there only minor slips?
34. This text makes effective use of linguistic modality.	propositional precision/grammatical accuracy	Can the writer make effective use of linguistic modality to signal the strength of a claim, an argument or a position? Are certainty/uncertainty, belief/doubt, likelihood expressed in an effective way? Are **all** modal verbs used consistently well?

Appendix 16 Checklist: final version – 30 items

The logic of the statements is not punishment or reward but merely noticing the presence or absence of a feature.
The reference scale headings in the middle serve as a direct link to the CEFR (2018).
The concept check questions on the right serve to ease the decision-making process.
If the answer to all the questions in a cell is in the positive, allocate a 1 indicating that the target trait is present. If there is a negative answer, allocate a 0.

Statement	CEFR reference scale	Concept check questions
1. This text is **legible**, i.e. the reader doesn't have to guess what the writer is trying to say.	orthographic control	Can you read the text without having to re-read words? Is the text legible? Can you keep your role as reader?
2. This text follows the standard **layout** of an essay.	orthographic control	Does the text follow standard paragraphing conventions? Are there at least four visible paragraphs (intro, more than one body paragraph, conclusion)?
3. This text is **clear** and **concise**.	overall written production/planning	Does the text give a positive overall impression? Are there signs of planning so that the reader's work is easier? Can you keep your role as reader?
4. **Spelling** is consistently accurate.	orthographic control	Are there only **two or fewer** spelling mistakes?
5. **Punctuation** is consistently accurate.	orthographic control	Is the text free from punctuation errors?
6. This text is the required length as defined by the task.	TASK	Is the text within the acceptable bounds as defined by the task (in this case cca.250 words)?
7. This text displays **situational authenticity** and **self-disclosure**, i.e. the writer gives the necessary details for contextualisation, such as role, location, situation, domain, etc.	creative writing & correspondence	Does the writer explicitly state their role, location, situation, domain, etc.? Does the reader get detailed realistic information about the writer's context, situational underpinnings, setting, etc.?
8. This text could function as a real life [**essay**].	creative writing	Does the text clearly state (a) the reason [*why getting a university education is no longer a guarantee of success 2018 july*] and (b) the details specified by the rubric (for and against, conclusion)?
9. The writer can explain the background to the [**problem**].	pluricultural repertoire	Can the writer reflect upon cultural values and practices? Can the writer deal with ambiguity in cross-cultural communication?

		Are the writer's reactions expressed constructively and culturally appropriately?
10. The content elements required by the task instructions are elaborated in appropriate **detail**.	TASK/thematic development	Are all the content elements (see item 8) discussed? Are these elaborated beyond being merely mentioned? Are they discussed in at least **one paragraph** each?
11. There is little or no **irrelevant information** in this text.	TASK/thematic development	Is the text limited to information that is directly related to the topic as defined by the task rubric?
12. The text contains genuine ideas.	creative writing	Does the text contain well-structured and developed descriptions? Is the text imaginative? Is the text presenting unexpected content?
13. The writer can present **multiple points of view**.	creative writing	Can the writer evaluate problems and proposals? Does the text contain reasons and relevant examples? Is the writer offering recommendations? Are there any **inconsistencies** or **controversies** highlighted?
14. The writer can communicate complex ideas in a logical way.	thematic development	Is there logical coherence in the text? Is there clear progression throughout? Are the key ideas clearly expressed in **each** paragraph?
15. The writer adopts the level of **formality** adequate to the topic, agents, situation, domain, etc.	sociolinguistic appropriateness	Are sensitive topics handled with care? Is the writer tactful enough so that the reader is not offended? Is the level of formality adequate to the communicative context and its agents? Can the writer frame critical remarks or express strong disagreement diplomatically?
16. Each **paragraph** presents one distinct and unified idea.	coherence and cohesion	Are the paragraphs of approximately the same length? Does each paragraph focus on one idea? Is each **body paragraph** longer than a single sentence? If not, is it complex enough to realize a paragraph?
17. Each paragraph contains a topic sentence.	coherence and cohesion	Are **all** the topic sentences relevant and meaningful? Do the topic sentences make it clear why the points are important?
18. The paragraphs create a **logically structured text** that is easy for the reader to follow.	coherence and cohesion	Are **all** the paragraphs linked in meaning as the text unfolds? Is there a semantic link over and above the presentation of a list? Is there a thematic development leading the reader from the

		introduction through the details to the conclusion?
19. This text shows a variety of cohesive devices.	coherence and cohesion	Does the writer demonstrate the controlled use of connectors and organisational patterns **in each paragraph**?
20. The text is characterised by complex grammatical structures.	general linguistic range	Do complex structures (gerund/infinitive, non-finite verbs, reported speech, indirect questions, adverb phrases, modals, modals in the past, etc.) **characterise** the text?
21. This texts demonstrates the use of **tenses and aspects.**	general linguistic range	Can the writer indicate shifts in time? Is there more than one tense used? Are the tenses used consistently well? Are grammatical aspects used consistently well?
22. This text uses language to formulate thoughts precisely.	general linguistic range/flexibility	Can the writer express him/herself clearly, without having to restrict what he/she wants to say? Can the writer express **finer shades** of meaning?
23. This text shows full consistency in the use of proforms.	general linguistic range	Is referencing used consistently well? Is the **noun-pronoun** agreement used consistently well?
24. Grammatical or linguistic **errors** in this text are difficult to spot.	grammatical accuracy	Is the text free from grammatical errors that could lead to misunderstanding? Is the text free from other linguistic errors that could lead to misunderstanding?
25. The text demonstrates the use of advanced word order and varying sentence length.	flexibility	Can the writer make a positive impact on the intended audience by effectively varying style of expression?
26. The text shows a good control of uncommon lexical items.	flexibility/sociolinguistic appropriateness	Does the text show a natural and sophisticated use of lexical features?
27. This text uses **a range of discourse functions** in a meaningful way.	turntaking	Does the text contain stock phrases (**at least 2**) to preface the writer's remarks? Are these used in a meaningful way?
28. The writer can use the English **lexicon** to express the intended meaning instead of periphrases or non-existent terms.	vocabulary control	Are **all** the words existing English items? Are **all** the words used in correct meanings? Does the writer find the correct word instead of relying on periphrasis? Does the writer find the correct word instead of creating one? Does the text demonstrate that the writer does not have to

		restrict what he/she wants to say?
29. This text is characterised by the use of **collocations and idiomatic expressions**.	vocabulary control	Is the language of the text expressive? Is the text characterised by idiomaticity? Are the idioms used meaningfully? Are there only minor slips?
30. This text makes effective use of linguistic modality.	propositional precision/grammatical accuracy	Can the writer make effective use of linguistic modality to signal the strength of a claim, an argument or a position? Are certainty/uncertainty, belief/doubt, likelihood expressed in an effective way? Are **all** modal verbs used consistently well?