Katalin **Bukta**

Rating EFL Written Performance

VERSITA

Versita Discipline: Language, Literature

Published by Versita, Versita Ltd, 78 York Street, London W1H 1DP, Great Britain.

ISBN (paperback): 978-83-7656-077-9

ISBN (hardcover): 978-83-7656-078-6

ISBN (for electronic copy): 978-83-7656-079-3

Managing Editor: Anna Borowska

Language Editor: Carl Becker

Cover illustration: ©istockphoto.com/ngkaki

www.versita.com

Contents

Introduction

Performance assessment for measuring writing ability has been used both in first language (L1) and foreign and second language (L2) pedagogy for a long time. Although this direct way of language assessment allows candidates to demonstrate their writing skills, the rating of their performance needs careful consideration. Several distinctive variables interact in written performance assessment: the rater, the rating scale, the performance, the task and the candidate (McNamara, 1996). The present book focuses on raters from among these variables and investigates written performance assessment from the raters' perspective. Their rating processes are influenced by several characteristics and the decisions are borne in their minds. According to Purves, "No matter how extensive or thorough it [rater training] may be, the rating is still a perception, a subjective estimate of quality" (1992, p. 118). This "perception", as he puts it, has generated substantial research recently. As a result, new frameworks of the rating processes demonstrate the complexity of the thinking processes raters go through in their decision-making (Cumming, Kantor, & Powers, 2002; Lumley, 2000; 2002; Milanovic, Saville, & Shuhong, 1996; Wolfe, 1997). However, little is known about these processes. As Lumley summarises his findings of extensive research into rating processes, "Nevertheless, much still remains unclear about what raters do when they assess writing texts" (Lumley, 2002, p. 7). These are the main reasons why it is worth exploring raters' behaviour, more precisely, the interaction between the text, the assessment scale, and the raters.

Observation of raters' thought processes, similarly to the examination of other mental processes is not easy; most frequently verbal protocol analysis is employed (Ericsson & Simon, 1993; Gass & Mackey, 2007; Green, 1998) thus offering a new avenue for explorations and think-aloud method for data collection offers a challenging enterprise.

Research into written performance assessment is a multifaceted phenomenon involving the nature of communicative language ability in general and writing ability in particular. The differences and similarities between L1 and L2 writing need consideration, as L2 writing research originates from L1 research. As far

as testing language ability is concerned, theories of language testing serve as a basis for written performance assessment. In order to be able to collect relevant data for observation of raters' rating processes, verbal protocol analysis should be studied. The context of the research is also presented: L2 education in Hungary with special attention to assessment of writing ability. The research carried out is meant to contribute to a better understanding of direct written performance assessment not only in the Hungarian but in a wider context as well.

Overview of the Book

The book consists of two main parts: Part One is the review of relevant literature and Part Two comprises of a pilot study and the main study on raters' decision-making processes in rating L2 written performance.

Chapter One overviews the nature of writing ability in L1 and L2 and their similarities and differences. Then, the relationship between writing in L1 and L2 is looked at. Theoretical frameworks of communicative competence shed light on the linguistic aspect of language ability. Finally, the implications of theoretical issues of writing ability on L2 instruction are presented.

Chapter Two focuses on issues related to measurement of language ability with an emphasis on language performance assessment. The most characteristic common feature in language ability assessment is to establish "what" and "how", the former refers to language ability or the construct, the latter to the task, to the measurement instrument for eliciting language and the way performance is evaluated.

The next chapter, Chapter Three narrows the discussion of language ability measurement to written performance assessment and presents ways of assessing writing. The discussion of the nature of written performance assessment focuses on validity issues and on various aspects of task characteristics. Finally, a definition of audience and test-takers' characteristics is provided.

As the discussion of written performance assessment is not complete in the previous chapter, a whole chapter follows on rating scales, rater variables and frameworks of rating processes, as these issues stand in the centre of the main study. Thus, Chapter Four is about rating written performance and the main frameworks that have developed in different testing contexts to reveal what constitutes rating and how raters behave during rating.

The overview of the literature is complete with Chapter Five which presents verbal protocol analysis as the main data collection method used in the research. First, the main issues in L2 acquisition research are presented and then, the focus shifts on the nature of verbal protocol analysis. The chapter attempts to provide a detailed picture of each stage of verbal data collection and processing.

Part II starts with Chapter Six, a pilot study on raters' thinking processes. Five raters' decision-making was followed to reveal some observable features of their rating. The aim of the pilot study was to try out verbal protocol analysis as research methodology for tracing raters' decision-making processes.

Chapter Seven introduces the research design and research questions of the main study. After the participants, the research instruments are presented, the data collection procedures are detailed in which verbal protocol data processing and coding scheme development play a central role. Finally, raters' division into competent and proficient groups is introduced.

Chapter Eight investigates features of raters' rating patterns starting with their gender distribution and language use in verbal protocol and protocol length. Then, the observations of sequencing rating and the pre-scoring stage and some emerging patterns are introduced. Finally, the chapter introduces raters' foci: management, rating, reading strategies are discussed together with raters' own comments.

Chapter Nine examines the strategies raters employed in more detail and deals with management, rating and reading strategies one by one and raters own focus comments are also interpreted.

Chapter Ten considers the four rating criteria in the order they appear in the rating scale and examines what strategies raters turned to when rating the ten scripts. provides a detailed analysis of raters' interpretation of the rating aspects of task achievement, vocabulary, grammar and organisation. Discussion of raters' rating patterns follows by looking at raters' focus when dealing with each of the four rating criteria.

Chapter Eleven follows the rating processes for the weakest script (N2) and the top script (N6) and investigates how the two groups of raters arrived at a decision. In addition, it examines the score choices and raters' thinking during rating.

Raters' perception of the rating task is in the centre of Chapter Twelve and provides some insight into the feedback they gave on the course in ELT, the rater training and the rating task.

Finally, some general conclusions are drawn, the limitations of the research are acknowledged together with its implications and ideas for further research.

Illustrations in the book are labelled sequentially and can be identified by a number indicating the chapter they appear in and another number provides information on their sequence within a chapter.

Tables are illustrations that comprise quantitative data discussed in the relevant chapter and can be identified by number of the chapter they appear in and a number indicating their sequence within the chapter.

Figures are those illustrations that either demonstrate what is explained in the text and they comprise diagrams to show different tendencies in the quantitative data. They are labelled according to the chapter they are in.

Excerpts are text parts cited from raters' protocols, they are included in a box and the raters' identification is written in the top left corner above the box. Each text unit (TU) is numbered as it appears in the transcribed protocol.

PART I

An Overview of the Literature on L2 Writing and Assessment

The first part of the book intends to provide an overview of the relevant literature on writing and assessment from Chapter One to Chapter Five.

Chapter 1

Writing Ability and L2 Writing Instruction

Introduction

The way people interact has always been central to research, since the need to communicate effectively is crucial. Modern age is characterised by advanced technology, especially in communication, however, we cannot exist without written communication. Before discussing how writing ability can be assessed, we have to have a look at the features of writing ability, what characterises the ability to write in one's L1 and what skills are needed for being able to express ourselves in writing in a new language (L2). This chapter first deals with the nature of the writing skill in L1 and L2 focusing on elements constituting language ability. The terms foreign language and second language are used interchangeably from now on and are labelled as L2. Then, frameworks of communicative language ability are presented briefly to highlight the linguistic aspect of language competence. In order to be able to measure language ability we have to have a clear picture of its components and the relationships between them. Writing ability in L1 and L2 is compared through looking at the way skilled and unskilled writers compose texts. In addition, the influence of reading on writing processes and the role of L1 in L2 writing are presented. Writing ability is strongly related to education and writing skill should be taught; therefore, a discussion of writing instruction both in L1 and L2 is relevant. The role of writing in L2 instruction is important, as language proficiency develops using writing as a learning tool, especially in the case of adult language learners (Weissberg, 2000). Writing skills are transferable from L1 to L2 writing and the transfer takes place during learning (Wolff, 2000). The discussion focuses on the main characteristics of L2 writing instruction, and, finally, main issues of L2 writing instruction are introduced briefly.

1.1 The Writing Skill in L1 and L2

The ability to write is not innate. Compared to listening and speaking, people need to reach a certain level of cognitive development before they can acquire writing skills. The skill comprises of several elements; it is not sufficient to learn how to form letters on a piece of paper with the help of a pen or a pencil.

Virtually anybody who does not have a physical deficiency can learn to speak, but writing skill development is much more complex, as it is part of general literacy skills, involving reading comprehension as well. Although speaking and writing are closely related, writing means more than merely recording spoken language (Weissberg, 2000).

Proficiency in a language can be described by making a distinction between four language skills and dividing them up according to the channel of communication: two of them, speaking and listening are oral skills, and two, reading and writing are written skills. It is also possible to differentiate between productive and receptive skills according to the mode: the former two involve language production (speaking and writing) and the latter ones involve receptive skills (reading and listening). Both L1 and L2 writers have to have all these skills at their disposal in order to be able to make the right decisions when choosing from the language store. It follows that writing is a productive language skill, which has both similar and different features in one's L1 and L2 (Harmer, 1991, p. 15). However, this distinction according to the channel and mode seems to be vague for understanding the processes involved in language use (Bachman & Palmer, 1996, p. 75). Although mental processes of language production are not easy to follow, several models have been proposed to explain the process a writer goes through in L1 and L2 writing development. These models are introduced in the followings to look into the nature of writing ability in L1 and L2 and their relationship.

1.1.1 Theoretical Models of Written Text Production

In order to have a better insight into L2 written language production, it is worth considering the models that attempt to describe how L1 writers approach and then perform a writing task. The models examine different elements of the writing process, such as the task, its environment, the writer and the audience; as well as provide an explanation of the relationship between these elements and the way they interact with each other. Thus, they facilitate the understanding of the cognitive processes, the knowledge needed and other underlying factors are easier to understand (Weigle, 2002, p. 23). The theoretical models presented below shed light on the main cognitive processes that writers follow when producing texts.

The model proposed by Hayes and Flowers (1980) includes three main cognitive processes that play a role in written language production: planning, translating and reviewing. Planning consists of organising, goal setting and generating; these steps are closely related to the writer's knowledge of the topic and strategies which enable him to organise thoughts coherently. During the translating phase the writer retrieves relevant information from long-term memory and transforms it into language. Translating in the model means

retrieving thoughts from memory and turning them into language. Finally, the text is reviewed to improve it in the translation phase with the help of reading and editing skills. The order of these processes is not linear, the writer can go back to some stages to improve the text, but most probably the steps follow each other starting from goal setting through planning, production to revision. These three main cognitive processes in the Hayes and Flower (1980) writing model are closely related to the writer's long term memory, where the knowledge of the topic and the audience is stored, and there are writing plans from which the writer can choose. When the writer is assigned a task, the task environment contains the topic, the audience and motivating cues with the text produced so far. The model centres on the role of planning and revising, and less attention is paid to sub-processes. Nevertheless, the recognition of the recursiveness of the writing processes is important and has served as a basis for further research.

The model developed by Hayes (1996, cited in Weigle, 2002, pp. 24-25) discusses the processes from two aspects: the task environment and the writer. In the model the focus is rather on the individual, whose motivation, working and long-term memory, and cognitive processes are examined in detail. Task environment comprises physical and social factors, which are the people involved including audience and the text, and the medium of writing. The role motivation and affect play in the model is considerable and they are related to the cognitive processes of text interpretation, reflection and text production. It means that the individual's success in performance depends on motivational factors. The information about the task and topic is stored in long-term memory, whereas working memory stores three types of information: verbal, coded and conceptual. These interplay with cognitive and motivational processes at different stages of text production. The other important feature that is highly relevant to assessment and instruction is that there are several reading types involved in the writing process: reading to evaluate, reading source texts and reading instructions. It follows that if the writer does not comprehend written texts properly, he cannot perform writing tasks. Hayes discusses the relevance of several types of reading in written text production: reading instructions, sources and the text during revision are the most important that need consideration.

Bereiter and Scardamalia introduce a two-model approach to writing (Bereiter & Scardamalia, 1987). They investigate the writing process by examining the differences between skilled and unskilled writers. Moreover, they introduce the notions of "knowledge telling" and "knowledge transformation": the first means simply recording speech, whereas the second relates to composing new language. The two terms refer to the following differences in text production: knowledge telling is when language is recorded with no or little planning, translation or other text production processes, while knowledge transformation involves all cognitive processes discussed above to come up with a new text. The model builds on the assumption that while people can learn to speak it is not

evident that they can also learn to write. The main difference between unskilled and skilled writers is in the use of strategies: unskilled writers employ fewer and simpler strategies than expert writers. Unskilled writers spend less time on planning and they revise less than expert writers. These features bring writing and speaking closer to each other, but the difference is that while in speaking there is interaction, during which clarifications and amendments can be made, in writing it is not possible. Moreover, the strategies skilled and unskilled writers use show different features, as expert writers' strategies do not develop from the simpler strategies used by unskilled writers, but they employ different ones. The Bereiter and Scardamalia (1987) two-model approach has informative implications to writing instruction, as it makes a distinction between skilled and unskilled writers; however, the model does not provide sufficient guidance regarding how to turn unskilled writers into skilled ones.

1.1.2 Theoretical Frameworks of Communicative Competence

Theoretical models of written text production explain cognitive processes, but fail to give an account of the linguistic knowledge that is needed for language production. In order to be able to understand the details of language ability, we have to turn to the notion of communicative competence which was introduced by Hymes (1972) in the 1960s and discussed further by Canale (1983) among others. The Bachman and Palmer (1996) model approaches language ability from language testing perspective, whereas Celce-Murcia, Dörnyei and Thurrell (1995) focus on language teaching aspects. The different models attempt to shed light on the elements constituting the notion of communicative competence, and explain the relationship between them. They serve as a basis for better understanding both L2 instruction and testing (Katona, 1995).

Several attempts have been made to describe what it means to be able to use a language and what elements such knowledge consists of. Chomsky (1965, cited in Celcia-Murcia, Dörnyei, & Thurrell, 1995, p. 6) refers first to language constituting of "competence" meaning the knowledge of the rules of the language and "performance" which is the actual use of the language. This rather static perception of language knowledge is further refined by Hymes (1972), who redefined these two elements and added a new, sociolinguistic aspect to it. His model constitutes four elements: formal features, realisation, appropriacy and accepted use (Hymes, 1972, cited in Katona, 1995, p. 69). This model takes context into account and considers it an essential element in language knowledge.

The need to explain how language can be used for communication and to integrate it with other components of language ability resulted in the model developed by Canale and Swain (1980), which consists of three components:

grammatical, sociolinguistic, and strategic competences. Canale (1983) proposes a revised version of the framework and he relates it to both foreign language teaching and testing. First, he explains the difference between communicative competence and actual communication saying that knowledge of the language system and skills is needed for actual communication: in his model communicative competence means both the knowledge of the language and the skills or the ability to use it. Canale distinguishes four competence areas which interact with each other and are required in communication. Grammatical competence is the knowledge of the language code, sociolinguistic competence is appropriacy of meaning and form of language, discourse competence relates to cohesion and coherence, and strategic competence which a language user refers to when there is a breakdown in communication or when he wants to make the communication process more effective.

The framework is further refined and explained by Bachman (1990) who tries to show the processes that interact between the elements and the possible application of the framework to language testing. He uses a slightly different term and describes communicative language ability as an interaction of knowledge structures and language competence with strategic competence, psychophysiological mechanisms and the context of situation. Strategic competence is not directly related to language and is viewed as a capacity to operationalise components of language ability. Language competence is then explained as comprising two main competences, organisational and pragmatic, which are further divided into grammatical and textual competences, and illocutionary and sociolinguistic competences respectively. The framework developed in Bachman and Palmer's model (1996, p. 62) emphasises the relationship between language use and its testing to explain how language ability can be assessed when performing test tasks. The elements of language ability, topical and language knowledge together with personal characteristics interact with strategic competence and are influenced by affective schemata; all these interact with language use and test task. In addition, Bachman and Palmer reconsider strategic competence in Bachman's model and introduce metacognitive strategies to demonstrate that language knowledge and metacognitive strategies make the language user able to comprehend and produce language.

The model developed by Celce-Murcia, Dörnyei and Thurrell (1995) views communicative competence from a L2 teaching perspective. It derives from synthesis and in some respect is a redefinition of elements of previous models proposed by Canale, Canale and Swain, and Bachman and Palmer. According to the model, communicative competence consists of five main elements: discourse, linguistic, actional, sociocultural, and strategic competencies. Discourse competence comprises elements that are necessary to produce texts: cohesion, deixis, coherence, genre, and conversational structure. Linguistic

competence corresponds to descriptive linguistic aspect: syntax, morphology, lexicon, phonology, and orthography (the last two depending on the mode). Actional competence is derived from sociocultural competence in other models and is treated separately in this model and defined for oral communication. It is related to understanding intentions and responding to them appropriately. In addition, it contains the following two broad components: knowledge of language functions and speech acts. The former is further subdivided into interpersonal exchange, information, opinions, feelings, suasion, problems, and future scenarios. Sociocultural competence refers to the knowledge of expressing oneself appropriately applying the following factors: social contextual, stylistic appropriateness, cultural, and non-verbal communicative factors. Strategic competence is considered in a broader sense including not only strategies used in case of breakdown of communication, but for production, as well. Thus, strategic competence consists of avoidance or reduction, achievement or compensatory, stalling or time-gaining, self-monitoring, and interactional strategies. The proposed frameworks of written text production and communicative competence attempt to describe language ability and in the followings focus shifts on the writing ability element and on issues of writing skill in L1 and L2.

1.2 Writing Ability in L1

Writing in L1 is a skill closely related to formal education and it should be explicitly taught as opposed to speaking, for example. This is the skill less often used outside instructional settings (Weigle, 2002, p. 1). Theories in L1 composition that serve as a basis for writing instruction focus on four components of the composing process: (1) the writer, (2) the audience, (3) reality and truth, and (4) language. Johns (1990) looks at each of the four components from three approaches of text production: process approaches, interactive views and social constructionist views. According to the process approach, the writer goes through several processes of text production, which are either focusing on freedom of expression or on problem-solving. Interactive views emphasise the interaction of the writer and the audience. Social constructionist view perceives the written product as a way of social interaction, so it plays an essential part in community life. The audience is looked at by expressionists as a creation of the writer, while cognitivists address the reader and consider his needs similarly to the interactive and social constructionist views. Reality and truth are discussed from different points of view: they can reside in the writer, or can result from interaction with the audience or can be determined by society. The language component is considered to be the writer's own in the process view, the interactive view takes both the writer and the reader into account, and the social

constructionist view emphasises the role of the community in language use. There is an intention to highlight these views, as they have strong implications on L2 writing processes (Johns, 1993) and on teaching and testing L2 writing.

As mentioned above, studies in L1 composition often distinguish skilled and unskilled writers and examine the processes that writers of different backgrounds and age go through (Raimes, 1985). Skilled writers go through the processes of goal setting, planning, organising and revising, but these do not follow each other as a set sequence, writers go back to certain steps as a result of reconsidering what they have and the pattern of text production is different from writer to writer. However, unskilled writers do not use each step and if they do, they spend much less time with them. They focus on grammatical accuracy more than on the content of the text. This distinction is investigated in L2 writing context with the aim to find out about similarities and differences between skilled and unskilled L1 and L2 writers (Krapels, 1990; Sasaki, 2000; Zamel, 1983).

Apart from looking at the components of the composing process and comparing different writers, the role reading plays in the writing ability development needs consideration. Reading can serve as a model for writing, as recognizing patterns in a written text promotes cognitive processes for producing similar patterns (Krashen, 1984). There are three types of reading-writing relationships related to L1. According to the first model, structural similarities between reading and writing are in the centre. It is called a directional hypothesis and it means that there is an underlying common feature by which reading promotes writing and vice versa. The model works in one direction only, so skills either transfer from reading to writing or writing to reading. The second, nondirectional or interactional hypothesis claims that there is an underlying cognitive proficiency that is involved in both writing and reading and they develop together in interaction. The third, bidirectional hypothesis means there is a relationship between the two skills and it is complex: they improve together but the proportions of this development can differ (Eisterhold, 1990). Although these models offer different focus on the relationship between reading and writing in L1 with respect to the transfer of skills from one mode to another, they are highly relevant in L2 writing instruction as well.

1.3 Writing Ability in L2

Writing ability in L1 influences the development of writing ability in L2. However, a generalisation of the differences and similarities between writing in L1 and L2 is not simple. The ability to write in an L2 has been of interest from several aspects. The L2 writer has developed the ability in L1 writing before writing in L2, so an L2 writer is on a higher level in his cognitive development. In addition,

an L2 writer is at a certain level of language proficiency in L2, which influences the ability to write in that language. Research has been conducted to reveal the relationship between L1 and L2 writing, to understand the processes of composing in L2, and to compare skilled and unskilled writers (Krapels, 1990). Moreover, written texts as products of the writing process have been examined to reveal the structure of the L2 text, to find differences and similarities between individual writers, and to analyse errors that students make when writing in L2. As writing takes place in a certain context, the interaction among the different elements of the context also needs attention (Archibald & Jeffery, 2000). Thus, L2 writing research focuses on the following areas: (a) texts written in L2, (b) learners' writing processes and characteristics, (c) ways of evaluating L2 writing, and (d) social contexts in which L2 writing or learning occurs (Cumming & Riazi, 2000, p. 56).

Furthermore, the relationship between reading and writing skills plays a role not only in the L1 writing ability development, but in L2 writing as well. As we could see above, the L1 writing ability development is related to reading and there is a transfer of skills possible from one mode to another. A similar process of transfer characterises the relationship between L2 reading and writing skills. In addition, there is a transfer across L1 and L2 possible and a certain level of L2 proficiency is necessary for this transfer to happen (Eisterhold, 1990). Writing skills can be developed by reading extensively for one's own pleasure and this has further implications affecting L2 instruction. Krashen (1985, pp. 18-19) applies his Input Hypothesis to developing writing skills and says that writers can improve their skills in writing if they read more. He emphasises the importance of reading for pleasure, as according to him, if readers enjoy reading, they will read more and attend to the text more.

1.3.1 The Relationship Between L1 and L2 Writing Ability

The nature of both L1 and L2 and their relationship has attracted considerable attention: L2 writing research originates in L1 writing research; most studies are conducted using case studies and think-aloud procedure, which are applied to a limited number of subjects (Connor, 1999, p. 307). Krapels (1990) discusses the relationship between L1 and L2 writing and pinpoints at findings that both support and question the relationship. Nevertheless, as in L1 writing research, findings are contradictory in some cases due to the low number of participants and samples examined. On the one hand, the way L1 and L2 writers compose is similar; the difference seems to be that L1 and L2 writers are at a different level of linguistic ability. On the other hand, some research showed that writers do not use the same strategies when writing in L1 and L2; they, for example, plan and revise less (Campbell, 1990). The strategies used in composing in L1 and L2 are similar and they can be transferred from one language to another. However,

it is not absolutely clear how, if at all, transfer happens in every case (Krapels, 1990). Further on, comparisons of skilled and unskilled writers show that an unskilled L1 writer resembles an unskilled L2 writer and the opposite is also true: a skilled L1 writer is usually skilled in L2 writing. It follows that a writer who is not a competent L1 writer is not good at writing in L2 and the quality of written performance in L2 depends on the writer's composing competence not on linguistic competence. Raimes (1985) expresses her concerns in connection with defining a writer as "unskilled" and attempts to define what makes a writer unskilled in a L2. She concludes that although there are differences between the composing processes in L1 and L2, L2 writers use strategies similar to their L1 strategies and try to express themselves in writing as much as they can bearing in mind that the language is not their native one, but a language they are learning.

The relationship between the writing ability in L1 and L2 is undisputable, as writers compose similarly in L1 and L2. However, there is a qualitative difference between the two composing processes and it is the fact that when one writes in a L2, L1 is at his disposal. This feature affects the composing processes in L2, as writers tend to switch from one language to the other to some extent. This language switching has been investigated by Woodall (2002) who makes a distinction between translation and language switch. The former results from instruction and is required by the task, while the latter is generated by the mental processes during composing and is not instructed. The study focuses on three factors that may affect L2 writing: L2 proficiency, task difficulty and language group to which the writer's L1 belongs. Results show that L1 use when composing in L2 does not have the same pattern regarding L2 proficiency and task difficulty. The different amount of L1 use does not depend on the writer's L2 proficiency and the difficulty of the task. L2 writers switch between the languages for several reasons, for example, they are looking for appropriate words, or they are planning what to write or they are editing their texts. There is some positive effect identified as well: writers seem to control their language use and they use L1 as an aid when composing in L2. However, the effect of the writers' L1 on writing in L2 is not apparent and Woodall attributes this to the academic context in which he collected the data (Woodall, 2002, p.14).

Although examining the relationship between the composing processes in L1 and L2 is important, there are other areas that need attention. The cognitive model of writing ability emphasises the recursive feature of the processes involved; this feature can be traced in L2 writing as well. The writer goes backwards and forwards when writing and both L1 or L2 can be used. The process is more complex than the same process in L1, as the writer goes back to the already written text more often, uses translation, repair and further planning more intensively. Manchón, Roca de Larios and Murphy (2000) examined three subjects' composing behaviour in their L2 and concluded that the choice of language depends on the cognitive demand of the task; if the task

is more demanding, writers use L1 for planning, problem solving and they also translate parts of texts. Freidlander (1990) gives an account of language use when composing in L2 and concludes that L1 aids L2 writing in planning and text production. If tasks are more related to the native culture, writers use more L1 during the writing process, and the product is of a higher quality. Krapels (1990) pinpoints at L1 use in L2 writing which influences planning, vocabulary use and can also be task-related. Zimmermann (1990) introduces a model, which focuses on a specific element of the writing process: sentence formulation. According to him, several subprocesses interact in sentence production and they result in tentative forms before acceptance. L1 can be present as a tentative form in problem solving after planning, but its use is rare according to his study. The gravity of his model regarding L2 writing processes lies in emphasising that L2 writers spend more time formulating language using different subprocesses than L1 writers.

1.3.2 Skilled and Unskilled L2 Writers

Similarly to discussions of L1 writing processes, similarities and differences between skilled and unskilled L2 writers need attention, since gaining an insight into the processes of skilled and unskilled writers can help L2 instruction in general and L2 writing instruction in particular. Both skilled and unskilled writers go through the composing processes in a similar way, they explore and develop ideas similarly. However, unskilled writers concentrate more on surface level and they are more concerned with linguistic problems, as they suppose the teacher will focus on them more than on the content (Zamel, 1983). Zamel goes on with her observations of composing processes and concludes that both skilled and unskilled writers plan their writing but no distinction can be made between them, as they use the same strategies of either note-taking or just thinking about the topic without writing anything down.

The cyclical feature of the composing process seems to be evident for all writers. However, skilled writers focus more on content and modify their drafts bearing content in mind, while unskilled writers reread and process smaller chunks of text which results in getting lost as far as meaning is concerned. The biggest contrast between skilled and unskilled writers seems to be that whereas skilled writers perceive composing as a creative and idea generating process, unskilled writers attend to putting ideas successfully together. These findings have implication for instruction, as they show that writing in L2 is part of the language learning enterprise and it is a problem-solving activity during which written texts are produced and new knowledge is gained (Zamel, 1983).

Sasaki (2000) compared expert and novice writers' composing processes using multiple data sources, including written products, recording time and pauses, and non-disruptive think-aloud procedure. The expert writers were

professors of applied linguistics and novice writers were college freshmen. The comparison revealed that expert writers spend more time thinking before writing, they complete the task earlier and come up with a longer text than novice writers. Novice writers' L2 lower proficiency level hinders their composing processes, they frequently stop to translate thoughts into English and they break the writing process to generate new ideas. Sasaki also found that six months of instruction in process writing did not result in an expected increase in writing quality: although writers employed strategies that expert writers use, their low L2 proficiency still was in the way of improvement.

To sum up, writing ability in L2 is closely related to L1 writing ability, as L2 writers build on L1 writing skills. However, there are some differences as well, which are mostly modifications of the processes used in L1 or features related to L2 learning. The complexity of the processes involved shows that in describing the L2 writing ability not only cognitive, but individual, social and cultural factors should be considered as well. Research into L2 writing process needs some further investigation, as findings so far are sometimes contradictory, but it is apparent that the results have implications for both L2 writing instruction and testing.

1.4 L2 Writing Instruction

L2 writing ability development is closely related to L2 language instruction, which takes the features of L1 writing development and L2 language development into account. There are several factors that play a role in L2 writing ability development in the classroom. These factors include task-related factors that influence writing in the L2 and depend on the topic of the task, the culture of the writer, the audience and the context. Kroll (1990) examines the time factor related to context in L2 writing and concludes that although there are some differences between compositions written at home and in class under timed conditions, the composing processes and the quality of the products seem to be similar.

As described earlier, writing ability develops similarly in L1 and L2 with an important difference: the writer in L2 has to have a threshold in L2 for being able to write in L2. Djuginović (2006) conducted a study among students of two age-groups and concluded that affective factors influence foreign language writing depending on age and proficiency level: the younger and less proficient students seem to show stronger motivation towards learning the language. Her other conclusion is that affective factors are more important in the case of complex skills, and productive skills, including writing are highly complex in nature. It implies that foreign language teachers should pay attention to promoting positive affective factors in learners in order to improve more effective learning.

A discussion of the features of L1 and L2 writing suggests that both the differences and the similarities have strong implications for classroom use. Writing, especially in a foreign language, appears mostly in educational settings. Historically, writing instruction first focused on copying; than, students were taught how to change texts and they also studied literary examples to improve writing skills. Finally, creative writing was introduced in the L2 classrooms (Bárdos, 1988). In order to be able to understand what goes on in the classroom, we have to examine what students can achieve and how they can develop their writing ability in L2 classrooms. In addition, knowledge of learning and teaching processes is needed for designing tests to get information on students' progress and achievement in writing ability (Cumming & Riazi, 2000). Writing in the classroom has several purposes: most often students use it as a tool for learning to reinforce oral communication or give an account of their linguistic or topical knowledge. The activities performed in the classroom range from highly controlled to free exercises: students copy from the board, record new material, perform grammar practice, write homework, write tests, etc. These instructional purposes can fulfil personal goals, as it is in the case of note-taking and later reading the notes for learning purposes, which involves the writer as the only audience.

Apart from instructional uses of writing, another type of writing in the classroom involves text production to perform different writing tasks. Most often the produced texts are read by the teacher and one or more peers and these texts do not have real communicative purpose either. They mainly serve as means to check learners' progress in the learning process. In classroom settings students have to produce different text types: Cohen lists the following types and provides a brief explanation for each "*expository writing* – to explain or inform; *persuasive/expressive writing* – to convince; *narrative writing* – to relate a series of events, *descriptive writing* – to offer a sensory impression of an object or feeling; and *literary writing* – to create exemplary text (in the form of a novel, poem, ballad and so forth)" (emphases in original, Cohen, 1994b, pp. 304-305). Regardless of the many possible text types, the potential of writing as a means for developing language ability and cognitive skills does not seem to be fully utilised in the classroom. However, considerable attention has been paid to writing instruction recently and a change can be observed in the approaches towards writing instruction.

1.4.1 Controlled Composition Approach to Writing

For a long time writing was considered to be a solitary activity and students produced different texts as home assignment or classroom task following some instructions. Afterwards, the product was evaluated and students received feedback, which concentrated on structural and mechanical features of the

text. Silva (1990) in a summary of the history of L2 writing describes the approach as spoken language written down which views language learning as habit formation. The teacher is interested only in the formal, linguistic form of language production and comments on it. The product-oriented approach ignores cognitive processes the interaction involved in text production and focuses on the outcome. Thus, writing resembles the writing model proposed by Bereiter and Scardamalia (1980) who refer to such language production as "knowledge telling" which is based on comparing content knowledge and discourse knowledge to the assigned task. From an instructional point of view the use of the product-oriented approach does not promote "knowledge transforming" which, according to Bereiter and Scardamalia, characterises skilled writers who go through the stages of planning, transforming and revision (cited in Weigle, 2002, pp. 31-32). The notion of product-oriented instruction approach can be characterized by focusing on the input text, language structures and translation (Cumming & Riazi, 2000).

1.4.2 Process-Oriented Approach to Writing

Recently, teachers have realised the learning potential in writing skills and introduced the so-called process approach to writing instruction (Cohen, 1994b, p. 305). A shift from the product-oriented approach which ignores expression and thought moves writing ability closer to language production and takes cognitive abilities into account. Students are guided through several stages of language production before the final version is completed and receive constant feedback both from their peers and from teachers. The process provides them with opportunities to revise, reconsider and rewrite their texts, which is similar to the way written texts are produced in real-life contexts. Silva (1990) describes the approach starting from the realisation that L2 writing processes are related to those used in L1 written production and writing instruction should involve students' thinking and creativity. The teacher's role in the composing process is important; it is her task to create a relaxed environment for students in which they can develop the appropriate strategies for written language production. Students have to get through steps of planning, revising, etc., which are similar to L1 written language production. Silva compares the description of approaches to L2 writing with drawing language instructors' attention towards the elements of writing that have to be considered in any writing programme; it is the interplay between the writer, the text and the reader in a certain context. Tsui (1996) presents her experiences related to the process approach in teaching writing. She concludes that although the process approach seems to activate students' language production skills to a greater extent, the product approach has some advantages and she proposes an integrated approach, which takes both teachers' and students' needs into account.

1.4.3 The Role of Writing in Communicative Language Teaching

The process approach to writing ability development has received some criticism, as it focuses on processes involved in writing, which may be different from individual to individual, who accomplish different tasks in different situations. For instance, the process approach to writing does not promote academic work, as the genres characteristic of study situations are not covered (Silva, 1990). Campbell (1990) points at the necessity of providing background texts in academic contexts and argues for their relevance in developing academic writing skills. She compared native and non-native students' academic writing using background texts. She concluded that non-native students benefit from reading and analysing authentic texts and they gain more confidence in using the style and the genres required in academic settings.

As discussed above, writing ability development in L2 is closely related to L2 language instruction. There are several approaches to learning and teaching foreign languages, all of which include writing ability development. Considering the theoretical frameworks of the communicative language competence, foreign language teaching has been designed along the lines of communicative competence. The approach highlights the different purposes writing can be used for in communication: it can be used for indicating actions, providing information, as well as for entertainment. The above mentioned characteristics, such as the differences between skilled and unskilled writers, the distinction between reading and writing, and the differences between approaches to writing have to be accounted for in communicative language teaching (Nunan, 1991, pp. 85-99).

1.5 Conclusion

The ability to produce written texts both in L1 and L2 appears to be complex from both theoretical and historical perspectives. The processes in L1 and L2 written language production show similarities and differences as far as the composing processes are concerned. One of the main differences between them seems to be that in L2 writing ability the level of L2 knowledge plays a role and is under continuous development. In this chapter an attempt was made to describe the theoretical basis of the nature of the writing ability. Then, the notion of communicative competence and the models that look at language competence from different aspects were presented. The linguistic aspect of language production in general and writing ability in particular was also introduced. Writing ability in L2 develops on the grounds of L1 writing, so in order to be able to explain how L2 writing ability develops, a comparison of the

nature of L1 and L2 writing is needed. Finally, the chapter dealt with L2 writing instruction and introduced main approaches to teaching writing. Writing in L2 classrooms has two broad purposes: instructional and composing purposes. Instruction related to written text production is characterised by product- and process-oriented approach. Communicative language teaching aims to focus on writing as a means for communication.

Language instruction is related to assessment, as we have to provide feedback to our learners and ourselves on the development of language ability. The following chapter discusses the nature of language ability assessment.

Chapter 2

Assessing Language Ability

Introduction

Chapter One provided an overview of the main theoretical considerations on the nature of writing ability, communicative language competence; writing ability in L1 and L2 are compared. The instructional aspect is presented with regard to the nature and ways of writing ability development.

In Chapter Two the aim is to list and evaluate the ways of making inferences about the degree and quality of this development and discuss what kind of measurement should be used. The type of measurement instrument depends on the purpose we want to use it for: the purpose may be evaluation without quantifying performance, or it may aim at ranking students, etc. Measurement is mostly carried out with language tests, which are used for two main purposes: to make inferences about learners' language ability in non-test situations and to make different decisions on the basis of test scores (Bachman, 1991, p. 680). Measurement can be used to define students' progress, to diagnose problems, to place the learners into different groups, or to collect data for research purposes. In addition, language programmes and different institutions can be evaluated with tests (Alderson, Clapham, & Wall, 1995).

This chapter focuses on discussion concerning fundamental issues of language ability assessment in general and language performance assessment in particular. First, a short historical overview of language testing is followed by some recent trends to give an insight into the development of language testing research and its objectives. Language ability and language performance testing are introduced characterising each component of a language test. To be able to design an appropriate measurement instrument or choose from the existing ones, the test method and test task characteristics should be considered. Discussions of scoring methods highlight the use of rating scales in performance assessment and deal with rater variables as well. Finally, ways of test score interpretation, test-taker characteristics, and test construction principles are detailed.

2.1 Assessing Language Ability

Assessment of L2 language ability has been in the focus of attention for a long time. On the one hand, it can have practical implications; test results

provide information for placement, selection and achievement judgements for educationalists and the job market as well (Alderson, 2001b; Bachman, 2000). Achievement on a test is informative for a student or a potential employee, which means that the test result should show test takers' true ability in the field tested. On the other hand, in order to be able to assign appropriate tests, research should be carried out into language assessment. Assessment of language ability can be conducted not only with tests, but by employing alternative ways, such as self- and peer-assessment, portfolios, teacher observations, etc., considering the purpose (Norris, 2000). Recently focus from traditional tests has shifted to performance assessment in the area of language testing, which originates in assessing other mental or physical abilities (Dietel, Herman, & Knuth, 2005). No matter how difficult it is to measure performance in general and language ability in particular, as it is the case with other mental skills, thorough research into the area helps to develop the most relevant measurement instrument and to employ the most appropriate one.

Before talking about different issues related to language ability measurement, the terms that are often used interchangeably should be defined. The term "assessment" is a superordinate term and usually refers to any type of measurement (ALTE, 1998, p. 135; Leung & Lewkowicz, 2006, p. 212). The term "test" refers to one type of assessment and "is a measurement instrument designed to elicit a specific sample of an individual's behaviour" (Bachman, 1990, p. 20). In addition, ALTE glossary (1998, p. 166) provides three definitions for the term "test": first, it is a synonym for "examination" meaning a set of components in assessment; second, a component of such procedure; and third, test is a relatively short assessment procedure. "Measurement" is usually applied for quantifying different characteristics, which may be related to individuals, processes or rules (Bachman, 1990, p. 18). "Evaluation" according to Bachman, is information collection for decision-making (1990, p. 22). These terms are used in this book interchangeably, except when their special feature is in the centre of attention.

2.1.1 History of Language Testing

The approach towards measuring language ability has changed during the history and has been dependent on and influenced by changes in sciences related to it, such as linguistics, psychology and education. The considerable advance in theory regarding language ability consisting of several components has resulted in new measurement instruments and statistical devices, and communicative approach to language teaching and testing has become more and more popular (Bachman, 1991; 2000).

Dörnyei (1988, pp. 17-32) identifies four stages in language testing history including the distinction of the three stages first proposed by Spolsky (1978,

cited in Dörnyei, 1988, p. 18) which he completes with a fourth one. The first stage, the pre-scientific period developed along the lines with the grammar-translation teaching method, which focuses on translation and structural accuracy and emphasises the authority of the teacher. This authoritarian approach affects testing and makes teachers responsible for assessment without any special knowledge or training for doing so and their judgements are accepted as valid and reliable.

The second, psychometric-structuralist period is characterised by a strong psychometric influence and resulted in the ultimate trust in discrete-point item testing, which focused on testing language elements separately. Language proficiency is measured as comprising four language skills (reading, writing, speaking and listening). In addition, there are phonology, morphology, syntax and lexis tests, as if each was a separate element of language ability as a whole. Consequently, the most popular form is multiple-choice test item type, which is analysable and which shows high reliability and validity characteristics. There is an ultimate trust in "norm-referenced" standard test, in which an individual's test results are compared to a "norm" group.

The third, integrative-sociolinguistic period centres on integrating testing techniques. The two most frequently used techniques are cloze tests and dictation. The items are based on authentic texts and resemble real-life language use. However, just like in the case of other item types, such as editing and white-noise test reliability and validity are questionable.

The fourth stage, the direct proficiency testing period developed on the grounds of the latest findings of communicative language competence theory. This stage is also referred to as the "communicative" language testing period. Communicative language tests have four main characteristics, which are closely related to language instruction: tasks contain an "information gap", they depend on each other and are integrated within a domain. In addition, communicative tests measure broader language ability components than their predecessors (Bachman, 1991, p. 675). The main principle behind these tests is to sample language performance adequately and measure this behaviour appropriately.

Dörnyei (1988, p. 32) mentions three components that need special attention in direct testing: what language behaviour is tested, how the rating scale is devised and how raters or examiners assess performance. This stage in language testing is called "naturalistic-ethical" elsewhere following Canale (1983) who emphasises the need for measuring communicative competence with tasks that require natural language use and the responsibility of test users in defining what to test as well as how and why (Chapelle & Douglas, 1993). The brief historical overview shows a variety of test types and stages of advance in testing language ability.

The test types above are still used with alterations in accordance with the latest research findings in the field of language testing and applied linguistics

aiming at objective, reliable and valid measurement of the construct in question. Research into comparing different test types to provide evidence for measuring the same construct shows that it is possible to assess language ability with tests that represent different approaches to language ability. Results of a large comparability study show that although design principles of Cambridge First Certificate in English test and Test of English as a Foreign Language are different, they measure the same aspects of language ability (Bachman, Davidson, & Foulkes, 1993). On the one hand, these findings provide evidence for the necessity of defining the construct carefully and only then it is possible to assign the test type to assess the ability in question. On the other hand, the need for further research in language testing is justified, which should cover different aspects of language testing.

2.1.2 Recent Developments in Language Testing Research

The 1990s saw a culmination in language testing issues: several meetings, studies, and published books mark the importance of the topic. The foundation of the International Language Testing Association in 1992, a revision of standardised international language tests and appearance of new ones are also remarkable contributions to language testing. Theoretical developments focus on rethinking the models of language ability, as it is inevitable to have a clear concept of the construct before deciding how to test it. In addition, further research into the issues related to different aspects of reliability and validity has been conducted; L2 acquisition experts and language testers have approximated their research closer to each other (Douglas & Chapelle, 1993). Regarding methodological development during this period, skill-based testing and developing tests for specific purposes have come in the foreground (Douglas, 1995). Growing interest in criterion-referenced tests has resulted in more intensive research into the area. Issues related to performance assessment have attracted researchers' attention (Alderson, 1999; Alderson & Banerjee, 2001).

Testers have reinterpreted traditional translation, dictation, and cloze test types. Application of new technology, such as computers and video has attracted researchers' attention in test design. They have developed new tests and have compiled standards of good testing practice. The need for having standards in language testing and teaching alike has been supported by the publication of *Common European Framework of Reference for Languages* (CEFR, 2001), which attempts to establish a common ground for teaching, learning and testing modern languages in Europe (Hamp-Lyons, 2004). The framework is not prescriptive; instead, it shows directions and raises questions rather than gives instructions for L2 education and assessment.

Alderson and Banerjee (2001; 2002) provide a thorough description of the state of the art at the beginning of 2000s. The first part of their review article

discusses research in self-assessment, testing young learners and alternative assessment. The second part deals with research into language construct validity and advance in assessing reading, listening, grammar and vocabulary, speaking and writing. The areas of interest in the testing profession presented above are supplemented with issues that need further investigation. They mention washback effect, ethical questions, politics, standards, English for specific purposes among others. Alderson and Banerjee raise the issues related to reforming national test design, which is supported with examples of the Hungarian Examination Reform (Alderson & Banerjee, 2001, p. 221).

Until recently, writing ability assessment mainly focused on "what" and "how" with the measurement instrument in the centre of attention. Research into assessment of written language ability has shifted from indirect to direct assessment of writing also referred to as performance assessment. Each element of the testing process needs thorough further investigation: the prompt, the rater, the rating scale, the rating process and the test taker have been and should be carefully researched. Concern about the characteristics of the rating process and rater behaviour in assessment of productive skills has generated research and has raised further questions to better understand written performance (Alderson & Banerjee, 2002).

Interest in alternative assessment has come to the foreground, as it is considered to provide a more complex picture of language ability. Studies discussing values of assessing language ability through work samples and the role of formative assessment in the classroom show another trend in the development of measurement (Leung & Lewkowicz, 2006, p. 226). Portfolio has been introduced as an alternative way for assessing writing ability; however, it raises concerns about reliability (Alderson & Banerje, 2002). The topic of alternative assessment is beyond the scope of the present book and it is not detailed here.

To sum up, as we can see, substantial contributions to language testing research have focused recently on performance assessment including written performance assessment and further research can be justified to this line of research by examining raters' behaviour and rating processes.

2.2 L2 Ability Assessment

According to the ALTE glossary, ability is "a mental trait or the capacity to do something" (1998, p. 134). As discussed in Chapter One, Bachman uses the term "ability" instead of "proficiency" and defines it as "being able to do something" (1990, p. 19). The framework for describing language ability presented in Bachman (1990) constitutes the "what" in assessing language ability and is based on theories of language ability. Now we turn to the discussion of features

of "how" to find the most appropriate means for assessing language knowledge. The measurement instrument, which is a language test in this case, elicits a certain sample of language to make inferences about overall language ability (Bachman, 1990, p. 20). That is why it is important to choose a sample carefully, which is relevant to the measurement goals. Language testing is different from measuring other mental abilities, as a language test uses language to make inferences about language ability: the object and the instrument of measurement are the same. To design the best instrument for elicitation, or choose from existing tests and use them appropriately is a responsible job. First, the difference between test tasks and non-test tasks should be considered, as testing is based on performance on several different tasks. The two types differ in their purpose: a test task is for measurement purposes, a non-test task is not. Moreover, if non-test tasks are used in language education, they promote language learning; it means that tasks for learning purposes are not always appropriate to measure language ability. Another distinction between a test task and a non-test task is in the context where it is used: we use test tasks to elicit performance in a testing situation, and we perform different tasks in real life.

2.2.1 Language Ability and Language Performance

The definition provided in the ALTE glossary for language performance is "the act of producing language by speaking or writing. Performance, in terms of the language actually produced by people, is often contrasted with competence, which is the underlying knowledge of a language" (ALTE, 1998, p. 156). Language produced while performing a writing or speaking task represents the person's underlying language ability and can be observed. The communicative era in L2 teaching and testing has brought up the notion of language performance. Performance assessment has a long history in different occupations to find the most appropriate person for a particular job. Assessment focuses on how well candidates can demonstrate their skills in a given profession. Assigning certain tasks to infer on one's overall ability makes the use of performance assessment justified in L2 testing contexts (Brown, 2004).

McNamara (1996) gives a historical overview of different models of language ability to explain their relevance to language performance and L2 performance assessment. He concludes that each model starting with Chomsky's (1965, cited in McNamara, 1996, p. 55) revolutionary distinction between "competence" and "performance" contains the notion of actual use, but the term "performance" is used to indicate several notions. In Hymes' (1972) model the ability for use and knowledge elements form communicative competence and there is a distinction between actual performance and potential for performance. The Canale and Swain (1980) model does not contain the potential of ability for use; they define it separately and call it communicative performance. The model

Bachman (1990) proposes for communicative language use treats strategic competence separately. The Bachman and Palmer model (1996) refines it further, as they replace the term "strategic competence" with "metacognitive strategies" to explain the differences in language users' approach to different tasks. In the light of the models of language ability, McNamara states that "a focus on the medium of the performance – language – as manifested under conditions of performance (as part of an act of communication) is characteristic of much L2 performance assessment" (McNamara, 1996, p. 45). He refers to two traditions that performance assessment builds on, one is the so-called "work sample" approach which originates from its non-language use. The other is the "cognitive" approach and draws on psycholinguistic processes. He also makes a distinction between "weak" and "strong" senses of L2 performance assessment. The weak sense means focusing on language use and the strong sense involves task completion using language. He argues for using performance assessment in L2 testing, as it resembles communicative language use more than traditional tests. Performance assessment constitutes the observation of behaviour during performing a task or tasks and assessors use a measurement instrument to make judgements of certain qualities.

2.2.2 Characteristics of Performance Assessment

The main characteristic feature of performance assessment is that candidates' behaviour is observed during carrying out a task, which is similar to real-world task completion. In broad terms, performance assessment requires students to use their knowledge and skills not only to complete a task in a certain school subject, but to act in a situation that resembles real life (Brualdi, 1998). That is why this testing method was initially used in vocational settings to measure candidates' ability for performing a certain job. Relevance of performance assessment in measuring L2 ability is apparent, as in order to make inferences on candidates' language ability, we can observe them performing on a task or tasks that model real-life language use. Thus, features of the test method are different from those of traditional testing, in which candidates' ability is measured with an instrument in order to arrive at a score. In comparison, in a performance test candidates' ability is elicited with an instrument to arrive at a performance which raters assess using a rating scale. There is an interplay between the elements of performance, raters develop their own rating processes and give a score on the basis of an agreed scale to assess the elicited performance. The differences between traditional and performance assessment characteristics are apparent when considering how McNamara perceives the interplay between the elements, as shown in Figure 2.1.

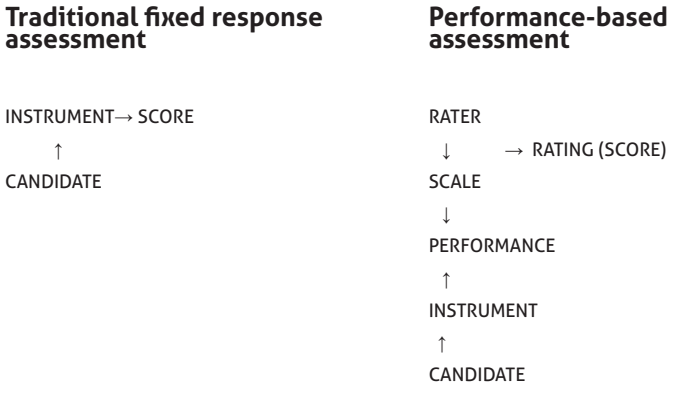


Figure 2 1. The characteristics of performance assessment (McNamara, 1996. p. 9)

Chapter Three provides further details of written performance assessment in which this interplay has a role and Chapter Four presents frameworks of rating processes to shed more light on the characteristics.

Language tests are supposed to measure language ability and according to the scores we characterise candidates' overall language ability. To achieve this goal, language tests should reflect real language use as much as possible. Sampling real-life language use is not easy; especially in case of general tests language use is so varied that it is not possible to model all areas of use. That is why the definition of the language use domain is important in a testing situation and the testing method should reflect domain characteristics. It also means that characteristics of language use task, situation, and language user should be similar to the characteristics of the language test task, testing situation, and test-taker (Bachman & Palmer, 1996, pp. 10-12). Comparison of real life and test performance characteristics is possible only if there is a clear definition of the construct that is the language ability we would like to make inferences about using tests. The following sections discuss the main features of measurement instruments.

2.3 Test Features

2.3.1 Test Purpose

As mentioned earlier, inferences about test takers' language ability are necessary prior to making several decisions. Regarding test purpose we can distinguish between large- and small-scale testing situations; the former

usually involves a bigger sample and is high-stakes in nature. The latter is often for classroom use in language instruction to make inferences about students' advance or achievement in the target language for instructional purposes. Tests can be categorised according to their purpose as placement tests, which are administered usually at the beginning of a course to assess candidates' language ability and assign them into different groups bearing in mind their level. Progress or achievement tests have a similar goal: to assess students' learning during a course and at the end of a course or a unit of study. Proficiency tests, on the other hand, aim to establish students' proficiency level with different learning backgrounds; whereas diagnostic tests aim at identifying students' weaknesses. Although these test types are different in their purpose, there are some overlaps, as, for example, an achievement test can be used for diagnostic purposes as well (Alderson, Clapham, & Wall, 1995, pp. 11-12).

The score on a language test reflects the language ability of the candidate with reference to a group performance or to a criterion level. The main characteristic feature of the former, the norm-referenced test is normal distribution to show the differences between individuals compared to the norm, or the reference group. A standardised test is norm-referenced and has three main characteristics: it is based on a standard content, administration and scoring are standard, and it is tried out. The so-called criterion-referenced tests are based on definition of the level of ability or domain and performances are compared to it (Bachman, 1990, pp. 72-76).

2.3.2 Testing Methods

Tests can be indirect and direct depending on the way they elicit language samples. Indirect way of testing involves measurement of isolated areas of the ability in question and testers make inferences on the basis of scores on its elements. An example of indirect testing involves making inferences on writing ability on the basis of a test on punctuation. Direct testing measures the language ability in question as it appears in real life situations. However, as it is not possible to observe language performance in real life, all tests are indirect in a sense and it means that the relationship between test task and real-life performance should be established with care in order to ensure validity of a test (Bachman, 1990, pp. 32-33).

According to the focus of a test we can distinguish between discrete-point or analytical and integrated tests. The former one focuses on separate elements of ability and intend to test them in isolation, whereas the latter focuses on the performance as consisting of several components (Alderson, 2000, pp. 206-207; Bachman, 1990, p. 34).

Elicitation of a language sample can be implemented by using different methods or techniques which should be chosen according to the test purpose.

There are several item types to choose from and the choice should be made carefully, as item types affect performance on the test, which is the so-called test method effect. If, for example, the candidate is not familiar with the item type or it is not clear what is required from him, the elicited performance will not reflect his ability (Alderson et al., 1995, p. 44).

There are two broad item types: objective and subjective, the former can also be called recognition, selected or fixed response type, as the candidate has to choose from several options. Objective item types, such as true or false, multiple-choice, matching, ordering, editing, gap-filling, one-word answer, etc. have several variations. Subjective item types, which are also known as constructed or extended response items require language production from the candidate; the response can range from one word answer to an extended elaboration of a topic. There is a wide variety of subjective item types, for example, compositions, essays, oral interviews, and information-gap activities. These item types require assessors' judgement on the quality of performance either on their own or involving more judges or assessors and usually they have a rating scale with certain criteria for correctness to compare the performance to (Alderson et al., 1995, pp. 46-62). Items are organised into tasks in a test and they include a rubric with instructions.

2.4 Qualities of Language Tests

As stated above, in order to make inferences about language ability, we have to measure it. This measurement can be a test, which should provide the user with useful information on language ability. Language tests should resemble authentic language use as much as possible, although sampling the language using the different criteria of authenticity is not easy (Chapelle & Douglas, 1993). Bachman and Palmer propose a model for test usefulness consisting of six qualities all of which should be considered equally and related to the purpose, target group and testing situation. Test usefulness comprises the following six qualities: reliability, validity, authenticity, interactiveness, impact and practicality (Bachman & Palmer, 1996, p. 17). These elements constitute the most important characteristics of language tests and need careful consideration when designing and implementing tests in a particular situation with a defined purpose for a certain group of test takers. However, they cannot be dealt with individually, they interact with each other and test designers should reach a balance when designing a test according to the actual testing situation.

2.4.1 Reliability

Reliability refers to consistency of measurement: test task characteristics should promote consistent measurement. Language test scores can be affected by test method characteristics, personal attributes which include both test taker and scorer characteristics, and there are some random or unexpected factors, as well. In order to maintain consistence of measurement and minimise measurement error we have to identify the sources of error and estimate their effect (Bachman, 1990, pp. 160-163). On the one hand, it means that if the test is administered at another time, the results should be the same. This internal consistency can be statistically measured by using different calculations, such as split halves or Kuder-Richardson formulas. Standardized large-scale tests are expected to have higher reliability coefficients (.80 or better) than classroom tests (at least .70) (Cohen, 1994b, p. 36-37).

Generalizability theory (G-theory) is another method to establish the reliability of a test (Lynch & McNamara, 1998). It is "a statistical model for investigating the relative effects of different sources of variance in test scores" (ALTE glossary, 1998, p. 146). First, the sources of variance are established, then, a study is conducted to estimate the sources of variance. On the basis of the results, the test is revised if necessary. It is followed by a decision study (D-study) conducted under operationalised conditions to estimate the variance of components (Bachman, 1990, pp. 187-188). Parkes (2000) emphasises the high costs of maintaining acceptable reliability of large-scale performance assessment. He says that in performance assessment there are more sources of measurement error than in traditional assessment, such as task and rater factors, which result in poor reliability and increasing the number of observations and tasks would raise the costs. He conducted two studies: one focused on scoring time and the other on financial consequences of a test which he included in his research. Parkes observed that it is possible to investigate cost-related factors in advance and it is possible to lower overall costs in a large-scale performance assessment test situation (2000, pp. 10-11).

On the other hand, consistency of rating has to be ensured and it means that raters should agree on their decisions not only among themselves (inter-rater reliability), but within themselves (intra-rater reliability) meaning that scoring the same performance several times should result in the same score. It is especially an important feature in case of subjective assessment type, as decisions depend on raters' behaviour: how consistently they can use the measurement instrument among themselves and when rating the same performance on different occasions. Written performance assessment involves text production on a certain task, which is then rated by one or more raters. It follows that reliability is influenced by task characteristics and variables of the scoring process (Weigle, 2002, p. 128-129). Raters' decisions influence the

reliability of written performance assessment, so finding out how they arrive at certain scores and what processes they follow can reveal how reliability can be maintained. In Part II of this book the study on rater behaviour attempts to shed light on rater reliability issues by examining their decision-making processes.

2.4.2 Validity

The concept of validity has been in the centre of attention since Messick (1989) pointed out its significance in language testing. Validity measure shows to what extent the test measures what it is intended to measure. It follows that a test cannot be valid if it is not reliable. In order to be able to make inferences about candidates' language ability on the basis of test scores we have to define what is to be measured. Interpretation of the test score is the justification of the evidence about the ability that the test intends to measure. Validity refers to the appropriateness of the measurement instrument and it shows how relevant the test is to the behaviour we intend to observe (Alderson et al., 1995, pp. 186-188).

For example, the question whether university students are placed in a language development course appropriately regarding their English proficiency level can be answered by assigning a test, which gives information on the students' academic language ability. Fox (2004) conducted a study to find out whether a high-stakes test at her university is a valid indicator of the language needed for academic studies. She compared teachers' misplacement judgements with students' scores on the test and concluded that although there are several factors that influence academic success, the higher the score on the test, the more likely the students' success is. In addition, test scores are used in the university programme to place students in English support classes according to the achieved score. Her results prove that the test has predictive validity and placing students in a support group relevant to their language needs has a positive effect on their academic studies, so early intervention is justified (Fox, 2004, p. 461).

Validity can be considered from different points of view: depending on the aspect of measurement we can distinguish between content, criterion-related, construct, systemic and face validity (Cohen, 1994b, p. 38). Content validity is the extent to which the test represents the knowledge area to be tested and therefore, the test is compared to a syllabus or a course material. It refers to a test feature to tell if the test requires test-takers to perform as in a real-life situation. A distinction can be made between content relevance and coverage; the former refers to specification of the language domain in question, the latter to the extent the test covers the areas of a domain to be measured (Bachman, 1990, pp. 244-247).

Criterion-related or criterion validity can be viewed from two aspects: concurrent criterion relatedness and predictive validity. Criterion here refers

to a criterion which is considered to measure the same ability. Concurrent ability measure shows the degree of agreement between test performance and recognized external ability. Predictive validity estimates the candidate's performance on a similar task some time in the future (Bachman, 1990, pp. 248-254).

Construct validity refers to the inferences we can make about the language ability in question based on the test results. Definition of the construct reflects the domain we want to measure and tasks should represent a sample that allows generalisations about the performance on target language tasks. It should be based on theory of nature of the ability it is intended to measure in order to be able to make informative judgements about an individual's language ability on the basis of scores achieved. As in the case of other mental abilities, language ability cannot be directly observed, we can only hypothesise about the relationship between a test score and the ability it is supposed to reflect (Bachman, 1990, pp. 256-258). As far as testing language ability is concerned, the test should sample how language is used and the observed performance should represent what candidates know (Leung & Lewkowicz, 2006, pp. 219-220).

In language performance assessment the validation process involves content analysis, empirical and theoretical investigations of both the tasks and the performance (McNamara, 1996). As far as written language performance assessment is concerned, the task should elicit the language we would like to measure, scoring should reflect the construct and the raters should apply the scoring criteria according to the construct definition (Weigle, 2002, p. 51).

Systemic validity refers to the effect that a test has on instruction and is referred to as washback including the consequences of tests on instruction, such as materials, student preparation and teaching methods (Cohen, 1994b, p. 41).

Face validity or test appeal is to show whether appearance of the test reflects what it intends to measure. Although this feature is not always considered as a rigorous characteristic as it is rather subjective, it contributes to perceiving the test being authentic and representing real-life language use (Bachman, 1990, p. 288).

2.4.3 Authenticity

Language test tasks should demonstrate how the language is used in performing target language tasks, and this correspondence is the authenticity of the test. Test task characteristics are compared to target language use tasks and the degree of authenticity is established. However, this comparison is not problem-free, "real-life" language use is difficult to define and people react to tasks differently and it is true for testing situations as well (Bachman, 1990). It follows that authenticity can be viewed from different aspects.

According to one approach, authenticity of a language test refers to the directness, meaning that language ability can be observed directly, without using any measurement instrument. Another definition approaches the notion by comparing test tasks to real-life use: tasks are authentic if they correspond to real-life tasks. The third approach defines authenticity in terms of face validity, which means that a test is authentic if it looks like a test (Bachman, 1991, pp. 689-690). In language performance testing modelling real-life can result in inauthentic test tasks, as a testing situation in itself is artificial. Moreover, test tasks are simplifications of natural ones and they only simulate real language use (McNamara, 1996). In a testing situation the audience is the examiner, who assesses the performance using a set of criteria. This artificial situation can affect the test taker, for example, in oral tests. In addition, the way raters interpret the criteria can also be inauthentic. Authenticity can be influenced by time factor as well, as in real-life task performance there is usually no restriction regarding the time needed for task completion. It is especially so in testing writing ability, as in a non-test situation usually there is no time limit for completing a writing task; moreover, if the context is academic writing, the student in a non-test situation has additional materials to refer to in a writing task (Weigle, 2002, p. 52).

A step forward in developing a more comprehensive definition of authenticity is proposed by Bachman (1991) who makes a distinction between situational and interactional authenticity which are closely related and share the same feature, both of them are relative. He says that a test task can be relatively low or high as far as situational authenticity and interactional authenticity are concerned. Situational authenticity refers to the correspondence of test method characteristics to real-life language use situations. It does not mean that the test tasks should be exactly the same as the ones in the target language, but they should have the same critical features (Bachman, 1991, pp. 690-691).

Communicative language teaching and testing is characterised by using authentic tasks in real-life situations with little or no mother tongue use. In addition, some national tests include translation or mediation task types for measuring L2 proficiency. These tasks are not considered to be communicative; however, as Heltai (2001) argues, translation or mediation activities form an integral part of L2 use; consequently they can be considered authentic.

No matter how test task authenticity is defined, the measure will always be relative, it depends on testers' and test-takers' perception of what they call authentic. Moreover, the issue of authenticity raises further problems; on the one hand, if a task in performance assessment aims at being authentic, it is fairly complex at the same time. On the other hand, if the task is complex due to the authenticity feature, the control over task difficulty is threatened. It follows that considering task authenticity issues cannot be divorced from dealing with construct validity, as authenticity is a characteristic that influences

the way language ability and assessment are treated (Leung & Lewkowicz, 2004, pp. 216-217).

The other type of authenticity is interactiveness which refers to the interaction between language ability, topical knowledge, strategic competence, and affective schemata with language test task characteristics. Test takers' language ability can be measured when performing on test tasks and this performance involves activating individual characteristics which interact with the test task. The individual characteristics are language ability, topical knowledge and affective schemata. This involvement is relative and depends on the extent of effort a particular task requires of the candidate (Bachman & Palmer, 1996, p. 25). When language performance is tested, not only linguistic knowledge is activated, the candidate goes through goal setting, assessment and planning stages to accomplish a task. It means that success in performing a certain task depends on the strategies used. As mentioned above, both situational and interactional authenticity are relative and we have to examine test taker characteristics together with target language tasks and test tasks to establish the relative authenticity and interactiveness of the test. It can happen that a task is highly authentic, but less interactive or highly interactive, but not authentic (Bachman, 1991, pp. 691-697; Bachman & Palmer, 1996, pp. 27-29).

The role of interactiveness is especially decisive in performance testing in general and testing speaking and writing in particular. In performance assessment of speaking there is interaction between the rater, the scale criteria, the performance, the task, the candidate and the interlocutor. It means that assessment largely depends on the way the interlocutor perceives the task, how he elicits language, or on assessor's expectations about the task (McNamara, 1997, pp. 453-454). McNamara concludes that performance assessment of speaking has social features and interaction between the components of the process influence the decision (1997). Similar features play a role in written performance assessment, which is in the centre of present study, as each rater has a certain background and individual features which influence the decisions they make about a piece writing.

2.4.4 Impact

Test design, administration, scores and other aspects related to testing have an impact at macro level, which means society, educational or even political systems, and at micro level, which means impact on participants including test-takers, test designers, authorities, examiners and other people who are influenced by it (Bachman & Palmer, 1996). High-stakes tests influence the macro level, for example, governments can use test results to allocate funding to schools and can have consequences on the quality of education. It follows that if schools need more financial support, they have to present higher test results, which

motivates teachers to enhance their teaching. Chapman and Snyder exemplify the impact of change in one of the high-stakes examinations with the intention to improve students' thinking and problem-solving skills and they introduced essay-writing tasks in a national examination. The change raised controversy in the society, as some people considered the tasks biased, teachers did not feel comfortable with the new task, and students and their parents thought the new tasks were unfair (Chapman & Snyder, 2000, pp. 5-6).

The impact of high-stakes tests can be viewed from the aspect of innovation and test scores can generate changes in educational systems. Wall (2000) distinguishes between innovation and change in her discussion of the impact of language testing on language education. She says that while change is not planned, innovation is a deliberate action and it means that it should be planned and prepared carefully (Wall, 2000, pp. 4-5). She lists her observations related to the issue and emphasises its complexity, the responsibility of participants and several other factors that have to be taken into account. Innovation affects individuals on three levels: changes of instruction content, the methodology employed and the attitude towards these changes. Individuals involved are teachers, test designers, educational policy makers, parents, stake-holders, and students, the potential test-takers (Wall, 2000).

2.4.5 Washback

Introduction of any innovation in education requires changes in teachers' work: they have to change their teaching material, their methodology, attitudes and behaviour. Such changes are neither automatic, nor easy; that is why teachers should be well-informed about the rationale of changes, they should understand how their practice needs to change and how they can implement them in their classroom practice (Fullan & Hargreaves, 1992).

Tests can generate different types of influence that have an impact on the candidates: before the test, which is the test preparation phase, during the test and after it. The type of feedback can define candidates' further advance, for example, or the test result can be used for selection, job application or promotion. Impact of testing on instruction is also called "washback" effect, which requires special attention and covers the relationship between testing and instruction as they influence each other considerably. The washback effect can be positive or negative: if the test encourages teachers to employ practices that promote learning in student preparation for the test, it means that washback effect is positive. On the other hand, if test takers' preparation needs practices that do not promote the development of language ability, the washback effect can be negative (Alderson, 1999). Empirical evidence though shows that this relationship is more complex in nature and a lot depends on teachers' attitude. Teachers have to be informed about the tests and prepared to integrate test

preparation in their teaching (DeVincenzi, 1995). Alderson and Hamp-Lyons (1996) followed two teachers' teaching practices to trace the differences between a test preparation course and a non-test preparation course. They conclude that the results are not as straightforward as they expected, and there are several effects that have to be taken into account when washback is examined. Findings show that test preparation affects the way teachers teach and the materials they use; therefore teachers should be prepared to harmonise their teaching with test preparation (Alderson & Hamp-Lyons, 1996, pp. 295-296).

However, washback is not always explicit; there are factors that have to be taken into account and thorough investigation and monitoring are needed before, during and after introducing any new test or examination. There has been some research into these factors to reveal more about the impact of testing on teaching. Wall (2000) lists the main areas of interest in this respect. She suggests a preliminary investigation of the antecedents to get a solid picture of the current situation and reasons for change. Such a "baseline" study has been carried out in several countries, including Hungary (Fekete, Major, & Nikolov, 1999), in which empirical data support the discussion of the necessity of a new national examination. A survey like this should involve all stakeholders in education: teachers, policy-makers, examination bodies, parents and students alike.

Bailey (1996) proposes a model in which she demonstrates the components and their interaction which is derived from the literature available so far and is built around three main components in the model. One component is the participants, who are learners, teachers, material writers, curriculum designers and researchers. Second, the processes and products that are related to learning, teaching, new materials and curricula and third, research results (Bailey, 1996, p. 264). Washback thus affects learners, their learning and test preparation, teachers' teaching methods and test preparation, materials and curriculum designers in their choices and researchers to find out more about the processes involved. Bailey makes several suggestions for promoting positive washback and areas for future research in the field. She concludes that all participants in testing and instruction should have clear ideas of test purpose, results should be informative, clear and fair. Tests should measure what the programme teaches based on clear goals and objectives justified by theoretical consideration. Finally, Bailey emphasises the need for students' involvement for positive washback, as self-assessment helps in learning and test-taking alike (Bailey, 1996, pp. 276-277).

2.4.6 Practicality

Practicality refers to test implementation, and is expressed by comparing available and required resources in terms of material and human resources, and time factor. Test design and administration consist of numerous activities, all

of which should be considered to decide about the practicality. Firstly, a test requires material resources, for example, rooms for administration, appropriate equipment and other equipment for both design and administration. Secondly, human resources should be considered, such as test administrators, raters, or interlocutors. Thirdly, test design, administration and scoring needs time, which should also be accounted for when planning a test. In performance assessment practicality plays an important role, as more raters are needed than in traditional testing situations. In testing speaking, for example, interlocutors are required to elicit the appropriate sample, and all of them should be adequately trained (Bachman & Palmer, 1996).

2.5 Test Method Characteristics

Test usefulness is realised through test tasks, and test task characteristics influence performance on a test, thus features of test method play a role in testing. Bachman (1990) and Bachman and Palmer (1996) provide a framework for test method characteristics, which consists of features of testing environment, test rubric, features of both input and output, and the relationship between the two. According to the framework, the following features of the testing environment should be considered: candidates' familiarity with the place and the equipment, their knowledge of the people involved in the test, time of the test, and physical conditions. Rubric provides information about the organisation of the test, time allowed for completing it, and features of instructions, such as language, channel, description of the tasks, and criteria for correctness.

Features of the input and response are similar as far as the nature of the language is concerned: length should be defined; elements of propositional content, organisation and pragmatic characteristics should be included. Format can be either aural or visual for both input and response. Format of input contains the mode, which is receptive, the form of input can be language and non-language or both, the language can be further characterised according to the vehicle of presentation, native, foreign or both; problem identification and speededness can also be included. Response appears in productive mode, the form can be language, non-language or both, and it can be either selected or constructed, in addition it can be native, target language or both.

Candidates in a testing situation are often faced with restrictions concerning the channel, the format of the language, and organisational, propositional, illocutionary characteristics can play a role. Time and length of response can also restrict the response. Relationship between the input and response can be reciprocal, as in a test of speaking, nonreciprocal as in the case of a writing test or adaptive when the order of the tasks assigned depends on candidate's

response. Although not each and every test task has all of these features, the framework serves as a basis for test design and evaluation.

Test method characteristics can be used to analyse test content to contribute to construct validity and test score interpretations. The information is necessary in both test design and understanding test results. Bachman, Davidson and Milanovic (1996) present a study on test content and communicative language ability component analysis of six reading papers of Cambridge First Certificate of English test. They devised rating scales for evaluating test method characteristics and communicative language ability components of input texts. Raters were asked to concentrate on the content of test items and ignore item difficulty judgements. The papers were used in operational testing situations, thus, achievement scores could be obtained and test content characteristics were compared to the performance. Outcomes of the research imply the necessity of devising a reliable measurement instrument for test method characteristics to be able to compare parallel and subsequent tests alike. In addition, Bachman et al. emphasise the importance of considering the effect that test method characteristics have on test performance (Bachman, Davidson, & Milanovic, 1996, p. 148). It is also useful to gain information on the test method for test designers, researchers and test-users to maintain the construct validity of tests.

2.6 Assessing the Four Language Skills

When talking about language tests, we usually refer to one or more of the four language skills, for instance, a test of reading comprehension. Practice of describing language ability, comprising four language skills of speaking, writing, reading and listening seems to be highly abstract in nature, though the distinction is present in both language instruction and testing. These skills are further grouped into two receptive (reading and listening) and two productive (speaking and writing) skills according to the mode of language production. Two of the skills, reading and writing are described as written skills, and two, speaking and listening as oral skills according to the channel of language. This distinction suggests the possibility of isolating language ability elements and dealing with them separately. Still, this categorisation is employed in both language instruction and testing for describing cognition and mode of language production (Weigle, 2002, pp. 14-15). Communicative approach to language testing also distinguishes between testing the four language skills; emphasis is on appropriateness when testing productive skills and on understanding when testing receptive skills, as Kitao and Kitao (1996) claim, adding that communicative features can be perceived as elements of a continuum. They say that a test can be communicative to a certain degree only, for example, it is not possible to speak without listening. In addition, there are further

substantial differences, such as, between listening to the radio or listening during communication. Language use is realised while performing on different language tasks and this categorisation is more useful for identifying the context of language use (Bachman & Palmer, 1996, pp. 75-76).

2.7 Scoring Methods

Depending on the test method, language is elicited using different item types and, then, candidate responses are assessed. Measurement usually results in a score, according to which inferences can be made about the candidate's language ability. Thus, in order to ensure reliable and valid score interpretations, there are several steps to follow: first, the construct or the language area to be tested should be defined theoretically, then, it should be defined operationally, and the method for quantifying responses should be developed. Candidate responses to the test tasks can be either selected or limited, or constructed as far as type is concerned. Selected or limited response type does not require judgement on part of the examiner and the scoring is objective, the examiner compares the response to the key provided by the test designer. However, in case of constructed response type the examiner's judgement is needed, the quality of the response is established by an examiner or more examiners, that is why it is subjective in nature (Alderson, et al., 1995, pp. 106-108). In case of a selected or limited response type the candidate has either to choose the answer from a number of options or provide a short, usually one word answer, and criteria for correctness means identifying what is considered correct and whether scoring is based on right/wrong distinction or it is possible to indicate degrees of correctness. Performance assessment is characterised by candidate's language production when performing a certain task and the response is compared to a descriptor or descriptors on a rating scale. For both objective and subjective types of assessment the criteria for correctness should be specified and the procedures of arriving at a score established (Bachman & Palmer, 1996, pp. 194-195).

2.7.1 Characteristics of Rating Scales

A rating scale in general is a scale that consists of ranked categories according to which subjective judgements are made. In language testing the different scales to assess language performance contain descriptors to make decisions clear (ALTE, 1998). Rating scales are mostly used for assessing language performance with so-called subjective item types, in which the language produced is compared to a scale descriptor for assessment. The most typical use of rating scales is in measuring speaking and writing ability directly. Scales are used in

many types of research; they are used to measure an individual's ability and have four main properties: they are distinctive, ordered, have equal intervals and have an absolute zero point. Thus, there are different scale types: nominal to identify categories, ordinal to refer to ranking entities and interval scales to tell the difference between entities. The fourth scale type, ratio scale is not widely used in language assessment, it is for finding out how many times one entity is different from another (Falus, 1996).

Scales have three main functions in measurement: they describe the level of performance, they guide assessors how to rate performance, and provide test designers with information on test specifications. Thus, there are three types of scales depending on who uses them and for what purpose: user-oriented for stake holders to have an idea of candidates' proficiency; assessor-oriented to aid rating; and constructor-oriented which reflects the test designer view of the construct (Alderson, 1991b, pp. 72-74). Pollitt and Murray (1995, p. 89) suggest a fourth one: diagnosis-oriented, which serves a diagnostic function and reports to teachers about candidates' performance. They find that Alderson's distinction lacks one function, namely reporting candidates' performance to non-assessor specialists.

Scales should reflect the underlying construct in a valid and reliable way. Moreover, scale descriptors should be comprehensible for all intended users. However, reliability and validity are threatened by several factors. In a classroom setting, although published standard rating scales are convenient to use, they can have some pitfalls. If the teacher uses the same scale on subsequent occasions, student improvement can result in different standards; descriptors can be interpreted differently; a published scale is too broad for the actual students; unwanted ceiling and floor effects can occur; and raters' standards can differ. Validity can be affected by teachers' objectives that are different from those of a scale; features in the scale do not occur in performances; and the scale can reflect different assumptions of language development (Upshur & Turner, 1995, pp.5-6).

According to their purpose, scales can be used for finding out about progress or achievement and can be used for comparisons which are either norm- or criterion-referenced. Scales have endpoints that should be clearly defined and the amount of information, brief or extended is also important in reporting scores. It is also important to know whether the scale has been evaluated empirically before operationalisation (Hudson, 2005). There are two scale types depending on their relationship to the task content: one is developed independently of content and context, the other is based on performance in a specific context. The former refers performance to the broad description of language ability, while the latter provides an opportunity to evaluate performance on individual tasks (Brindley, 1998).

Generally, there are two types of scales used in language performance assessment: a holistic scale to assess the performance as a whole, that is why this type of scale is called impressionistic, and an analytic scale which is to assess the performance according to its components (Alderson et al., 1995, pp. 108-109). Rating scales consist of numbers, letters or labels to describe performance; these are organized into bands to make a distinction between levels. Number of levels can be even or odd depending on whether an exact middle point in the scale is necessary or not. Number of bands influences reliability, theoretically the more bands, the more reliable the test is; however, it is difficult to handle too many bands with accompanying descriptors. Moreover, raters have to make clear distinctions between the scale points and if there are too many their attention can be distracted from the performance (Bachman, 1990, p. 36). The solution can be to have so-called "empty" bands, which do not have descriptors but make more detailed assessment possible if the performance does not match an upper scale point but exhausts descriptors for a lower scale point. Rating scales are sometimes divided not only into levels or bands, but the performance can be assessed from different aspects. For example, in written performance assessment task achievement, vocabulary, structures and organization can be assessed with the same weighing or assigning different scores for different aspects. Rating scales for written performance assessment are discussed in more detail in Chapter Four where the discussion is narrowed down to the focus of the present study.

Thus, a wide variety of rating scales can be produced depending on the purpose of the test and the area of language ability to be tested. In order to arrive at reliable and valid judgements, raters should understand the rationale of the scale developed and should be able to use them consistently throughout the rating procedure.

2.7.2 Rater Variables

Language performance is assessed by using a rating scale which involves assessors, who can be called judges, examiners or raters. Raters are involved in assessing speaking and writing skills directly on tasks to observe language performance using some kind of a rating scale. Raters play a decisive role in this type of assessment and their characteristics influence the score or scores assigned. Rater characteristics are sources of measurement error, which have to be kept at low level and monitored throughout the rating process. Measurement error occurs as a result of rater effects, which include the halo effect an impression of a performance which influences evaluation of other performance. Stereotyping is raters' expectations of a performance, their experience or background knowledge, severity or leniency, and not paying equal attention to all bands of the scale (Rudner, 1992).

There are two types of reliability related to rater variables: inter-rater reliability is the degree to which raters' decisions are similar, and intra-rater reliability, which is raters' consistency in their own decisions (Alderson et al., 1995, p. 129). In addition, raters should be familiar with the marking system which they should employ consistently and should be able to deal with unexpected difficulties (Alderson et al., 1995, p. 105). That is why rater training is inevitable, and it is especially so in case of high stakes large-scale testing format.

However, research has highlighted certain questions, such as the changes in rater judgements in time regardless thorough rater training. Congdon and McQueen give an account of a study on a large-scale test data set to investigate rater behaviour. They conclude that stability of rater severity is time dependent, it means that raters' judgements change with time compared to their own ratings and to each other as well. They claim that not only thorough rater training is needed at the beginning of a large-scale test, but continuous rater monitoring is also required to arrive at consistent scores (Congdon & McQueen, 2000, p. 175). Issues of rater training are discussed with special attention to training raters for written performance assessment in Chapter Four.

2.7.3 Score Interpretation

The measure that is closely related to score interpretation is content validity, which contributes to gathering information on candidate's language ability. The score should allow generalisations of the target language use (Bachman & Palmer, 1996, pp. 225-226). However, the score reflects not only language ability of the candidate, but is a complex indicator of different influences, such as strategy use and test-taker characteristics. That is why test score interpretation has to be considered with care and research into its characteristics has been carried out for the last decades (Bachman, 1991, p. 676-677). If test scores reflect language ability in non-test situation, they should provide substantial and comprehensive information about the candidate's language proficiency. Test scores are used for making decisions and thus they have considerable impact on different parties involved. This impact raises ethical issues, for example, individuals should be informed what their test scores are used for and how. Test scores provide information about a candidate's progress and quality of instruction at the same time. Impact of test scores on education is related to ethics, as the effect of test scores can influence instruction and educational policy as well (Leung & Lewkowicz, 2006, pp. 224-226).

It is not only test administrators who should be able to interpret test results, but it is important for the candidate to understand what the expectations are in order to prepare for the test. It is also inevitable for teachers to adjust their teaching and evaluate its effectiveness as reflected in students' scores. Language learners usually have an idea of their proficiency level and have

certain expectations towards tests. In case they do not understand what is expected they cannot prepare for the test appropriately. Teachers should raise students' awareness in connection with the test requirements and should make clear how they are related to language proficiency.

2.8 Test-Taker Characteristics

As tests elicit a sample of language ability on a particular occasion, test-takers' behaviour, mood at the time of the test and their attitude towards a testing situation determines the performance on the test to a large extent. Therefore, especially in case of high-stakes standardized tests, they have to be informed about the content of the test and strategies to use in accomplishing different tasks. Calkins, Montgomery and Santman (1999) compiled a list of activities that teachers can include in classes when preparing for a standard test. They highlight the importance of using different techniques to identify the main ideas, to concentrate on the question and understand what the task requires from them. Test-takers very often run out of time, they cannot complete all assigned tasks, because they cannot estimate the time needed for each task. Bukta (2000) conducted a survey of Hungarian students' test-taking strategies and found that students often work on a test paper in a linear sequence, without any planning of procedure. This can result in leaving the last questions unanswered, as many test-takers run out of time.

Test designers have to take test-takers' characteristics into consideration and design tasks which elicit the targeted performance sample if appropriate strategies are used. Cohen (1984) gives an account of the relationship between test constructors' and test-takers' intentions on the basis of verbal data obtained from test-takers when doing cloze and multiple-choice test tasks. If test-takers follow the task instructions carefully, the results on the test are better and the implication for teaching is that teachers should spend sufficient time preparing students for a test. In addition, test constructors have to state explicitly what test-takers are supposed to do and how they are expected to proceed on a test. Test-taker characteristics affect performance on tests, so test designers should consider them in language test construction (Bachman, 1991, p. 675).

2.9 Language Test Construction

Language test construction and administration require careful planning and professional judgements in large-scale and in classroom testing alike. There is no difference between designing a high-stakes and a classroom test, both of them should be compiled along the same lines (Alderson & Clapham, 1995).

There are three main stages that consist of several components. The first, the design stage involves description of the purpose of the test, identification and description of construct and tasks, test-takers and criteria for correctness. The second, the operationalization stage involves developing test tasks and a blueprint. The third, the administration stage is about trying out the test and collecting information for making inferences before operational use (Bachman & Palmer, 1996, pp. 86-93).

Language ability in question should be described very carefully in order to be able to design items that sample it appropriately. Alderson et al. (1995) propose a slightly different model for test design in which there is more emphasis on test specifications and test construction. It starts with stating the test purpose, and then, detailed test specifications are developed and guidelines for item writers and their training are compiled. It is followed by moderation and piloting tasks, which are analysed and revised. Examiner training is conducted and their reliability is monitored. Reporting scores and setting pass marks stage is followed by test validation and post-test reports. The last component of the model is test improvement.

However, the relationship between the test designer's intentions and the items is not always straightforward. Alderson (1993) argues that in deciding on the content validity it often happens that judges' opinions differ as far as the item content is concerned. It implies that content validity should be established using several means and careful pretesting is needed. Apart from these, there are other reasons for pretesting: it is not sufficient to establish item difficulty by only looking at the items. Similarly, it is not possible to provide well-justified judgements on the cut-off scores as well.

2.10 Conclusion

This chapter attempted to survey and discuss some main issues related to assessing L2 ability. The discussion started with a brief history of language testing to show the main milestones in development and relate language ability assessment to other fields of linguistics. Measurement of L2 ability has undergone substantial change which does not mean putting traditional methods aside, but building on research findings and revising existing methods. In addition, recognition of relevance of assessing people's ability in performing different jobs and assessing language performance has resulted in new approach to language testing. Language performance assessment attempts to model language use in real-life situations, which needs thorough consideration of the construct, or domain in question and of test characteristics. Inferences about L2 ability on the basis of performance test results depend on the interplay between several variables, which influence the outcomes. These variables include task

characteristics; a task should sample the language ability in question for making informative judgements about language ability.

Measurement is accomplished using some kind of rating scale which is used by raters or examiners. A rating scale contains descriptors organized in bands or levels and raters compare the observed behaviour to them to arrive at a score. Thus, raters' understanding of rating scale descriptors, their perception of task requirements and the observed behaviour are decisive components of the rating process. Measuring language performance of speaking and writing skills focuses on these issues. In the following part the attention from the general issues of language performance assessment shifts to assessment of written performance and an attempt is made to shed light on components that play a role in the assessment process.

Chapter 3

Assessing Writing Ability

Introduction

Discussions of theoretical issues related to writing ability in Chapter One was followed by an overview of the main trends in language assessment in Chapter Two with a special focus on language performance assessment. This chapter focuses on written performance assessment building on theoretical knowledge of language ability in general, and writing ability in particular relating it to measurement of writing ability both in L1 and L2 with special attention to written performance assessment. First, the history of writing assessment is presented briefly, and then, similarly to the theoretical discussion of writing ability and written language instruction which share some features in both L1 and L2, the main issues in L1 and L2 writing assessment are introduced.

The complexity of written performance assessment lies in the interaction between three main elements: test construction, in which the test designer's understanding of written performance is reflected; test-takers' understanding of task requirements and their writing ability; and raters' understanding of task requirements, the rating process and their perception of writing ability (Cohen, 1994b, pp. 307-308). Alternative ways of assessing writing ability are presented to provide a wider perspective for written performance assessment. Features of writing performance assessment are dealt with one by one without making an explicit distinction between L1 and L2 writing to provide a deeper insight into task and writer characteristics. As rating processes are in the centre of interest in the study presented in this book, a whole chapter, Chapter Four, is devoted to one aspect of assessment – rating written performance.

3.1 Written Performance Assessment

History of writing dates back to 12th century B.C. to China, when it was used to conflate knowledge through writing. In those days the way a written piece was produced was not important. For a long time ability to write was considered as a mark of level in education. In the 19th century written tests became widespread for evaluation in different institutions of education, as it was a tool for assessing knowledge. The role writing played in testing changed only in the 20th century.

Similarly to different models for describing writing ability, there are several ways of assessing it. There is a distinction between indirect and direct testing of writing; the former aims at assessing different components of language ability separately, the latter focuses on the ability as a whole. The ultimate trust in "objective" assessment in the 1950s and 1960s affected testing writing ability, as literacy was viewed as a composite of discrete skills that could be measured separately and objectively. Thus, in an indirect test of writing ability candidates are required to recognise components of writing and not to demonstrate their ability in composing a text. Direct assessment of writing ability came to the foreground in the 1970s and it involves language production. It is evaluated subjectively, using different measurement instruments: analytic, primary trait and holistic scales or impression marking is carried out (Hamp-Lyons, 1995a; 2002). Direct tests of writing ability have the following characteristics: they elicit a piece of continuous text of about 100 words; candidates are given instructions but still have freedom to express themselves; the text produced is assessed by one or more raters; rating is realised using rating scales which reflect the level of ability in question; the result can be expressed by a single score or more scores or with comments (Hamp-Lyons, 1992, pp. 5-6). There are several constraints that should be taken into account when assessing writing directly. The variables are related to the task, the test taker, the rater, the scoring process and even time which are detailed in the followings.

Alderson and Banerjee (2002) discuss the assessment of written performance as one of the skills in their review of the literature on testing up to the turn of the 20th century. It can be a mere coincidence that they place testing writing ability last in their summary and it would not be fair to conclude that writing is the least important among the skills measured in tests, but the case most probably is that testing written performance is very complex and by far not researched exhaustively. They raise several issues in connection with testing written performance and draw a parallel between testing speaking and writing. In addition, Archibald and Jeffery (2000) in their editorial to an edition of *Learning and Instruction* journal devoted exclusively to L2 acquisition and writing, emphasize the importance of researching assessment of written performance. They list several forms of assessment in different environments for a variety of reasons, including classroom assessment or linguistic evaluation of written products.

Although writing ability assessment in L2 has developed on the grounds of measuring writing ability in L1 there are several differences. While writing in one's native language, L1 language ability is supposed to be at a certain level, however, when writing in a L2 we have to take L2 language ability growth into account. The assessment of L2 writing, similarly to the assessment of other skills, can have different purposes: it can be small-scale for classroom applications, and large-scale to make inferences and decisions about a whole

population. The shift in approach to writing as a skill affected the assessment of L2 written performance significantly. It is not the question of being able to use vocabulary appropriately or manipulating structures correctly, but the ability to express oneself in L2 in writing. Thus, the construct of the performance has to be considered, what language is exactly required for what purpose and what audience should be borne in mind. It has affected substantially the way written performance is assessed, it is not enough to reflect on the language used any more, but the content should be considered first (Leki, 2004).

A test of writing in L2 is like a test of writing in L1, can be indirect to measure the knowledge of particular language areas, and direct, which involves producing written texts to present writing ability as it appears in real-life contexts. Depending on the purpose, different measurement instruments can be used: indirect test types are usually objectively scored, whereas direct tests of writing ability require subjective scoring in which the result depends on the interplay of several factors.

The way writing is taught influences how writing ability is tested. Cohen (1989) notes the difference between the process approach to teaching writing and the way it is tested. Writing ability is most often tested under timed conditions with no scope for revision and redrafting, candidates usually have sufficient time for brief consideration and then they have to produce the final version. Cumming (2001) gives an account of several writing instructors' assessment techniques, which are used to evaluate student achievement in both specific and general purpose courses. The interviewed teachers claim to use proficiency tests, especially for being able to report student achievement. They also use rating scales and competency assessment, the descriptors of which they explain to their students. These exams are considered to be typical for evaluation of students' progress on particular courses. Grading is accomplished in some cases using codes and leaves the students to improve the written piece and mostly positive features are mentioned in evaluation.

Recently in writing ability assessment considerable attention has been paid to performance assessment, which intends to measure writing ability directly. Similarly, alternative ways of assessment have been introduced. Portfolio is one of the alternative methods that have been used for years now with which continuous assessment is possible. Alternative ways have emerged to fulfil the need to assess writing ability as naturally as possible and still maintain acceptable reliability and validity.

3.2 Alternative Ways of Writing Ability Assessment

Writing ability is part of literacy skills and is closely related to instruction, which plays a significant role in teaching and testing all school subjects. Writing is

present in the classroom not only as a means of recording learning material, but students give an account of their learning in the form of written tests. First, as discussed above, the focus was on "objective" or indirect way of assessment in all areas including writing. Then, researchers objected to the prevailing use of "objective" measurement and they considered other ways of assessment. The introduction of direct tests raised the issue of reliability and validity in general and in writing assessment in particular. Some expressed dissatisfaction towards timed direct written tests and turned to portfolio assessment, which is more context-focused, but still not researched adequately. The other solution would be computer-based assessment, which has potential, although not sufficiently elaborated yet. L2 writing is sometimes mentioned together with "non-traditional" writers, who are people from different English speaking countries, with a variety of cultural background and language variety, which raises ethical problems in setting writing tasks appropriate to their way of thinking (Hamps-Lyons, 2002).

Chapman (1990) emphasises the need to employ an integrated approach to writing assessment and not to restrict it to writing classes but to assess students' writing ability in all subjects. She calls this approach authentic writing assessment and presents several examples for integrating writing ability assessment into testing different school subjects. She says that in a subject-related test it is possible to focus on students' subject knowledge and their writing ability at the same time.

Stern and Solomon (2006) argue for the importance of providing positive, focused feedback with pointing out the weaknesses and areas that need improvement. They researched written feedback of instructors of a whole university faculty for portfolio coursework to find out how the above mentioned three principles of providing positive, focused and informative feedback were applied by teachers of different disciplines. They identified 23 categories which were grouped into four main comment categories or levels. The first level comprises global comments; the second, so-called middle-level comments focused on forming ideas; the third, micro-level category covers comments on technicalities, and the fourth level contains other comments that do not fit any other category mentioned above. The comments are also categorised as positive, negative or neutral. They conclude that the ratio of meaningful comments needs improvement and these comments promote students' better understanding of the subject and learning. In addition, as a result of feedback students' academic writing skills improve parallel with their academic advance (Stern & Solomon, 2006, pp. 31-32).

The other approach to authentic writing assessment is related to impromptu features of test tasks. There is a difference between writing under timed and not timed conditions, the limitations of impromptu tasks, such as lack of sufficient time for composing, or the inadequacy of a single task to assess students'

writing ability are features that affect students' performance (Blattner, 1999, p. 232). This has turned attention towards alternative ways of written performance assessment, such as process approach and portfolio assessment.

Similarly to process- or product-oriented writing instruction, assessment of writing skills can be viewed from these two aspects. While in a product-oriented writing test the final draft is assessed, written under timed conditions, in a process-oriented one the focus is on the process of text production. Cho (2003) argues for a process-oriented approach as a more realistic way to assess academic writing ability in non-timed conditions. He says that as writing for academic purposes in real context is not bound to time limits, its testing should not be administered in timed conditions. As writing is a cognitive process involving several steps before the final draft is produced, its testing should take this into account. If the difference between good and poor writers is that better writers spend more time with revision and composing, it is justifiable to let candidates spend as much time on composing as they need. In order to find evidence for the differences between the two types of writing tests, he conducted a study to compare a timed traditional essay type writing examination and a workshop-based essay examination. His findings show that if candidates are given more time and opportunity to work on their written texts, as in a real situation, they produce higher quality essays. This approach and a variation of it, portfolio assessment, is highly authentic and has considerable learning potential but their reliability and validity are questionable.

Portfolio assessment has been used for assessment purposes for a long time in several disciplines. It is a collection of candidates' products which are collected according to strictly specified criteria. There are educational systems in which portfolios are used for quality control purposes to maintain accountability of an institution. It is through students' written performance that the quality of education is assessed. Portfolio, among other written products, can contain a reflective letter, which consolidates students' reflection on their progress in composing. The piece of writing is assessed bearing in mind holistic scoring and provides information on classroom instruction and it also shows the extent to which curriculum requirements are followed. Therefore, the piece of reflective writing in a portfolio serves as a benchmark for teachers to consider curricular expectations, it is informative to students about their development, and it validates the writing curriculum (Scott, 2005).

Relevance of assessing written performance using portfolio assessment originates in the way writing tasks are accomplished. As there is no time limit, candidates can spend as much time as they need; in addition they can refer to additional materials and produce several drafts as well (Hamp-Lyons, 1992, pp. 20-21). Hamp-Lyons says that written products included in portfolios are not usually written under timed test conditions and sometimes they are used to complement other assessment types. Although the potential of examining

candidates' performance from many aspects is justified, complexity and variety of written samples collected in portfolios raise validity and reliability issues (1992, pp. 27-28). Portfolio assessment is more widespread in classroom contexts than in high-stakes formal assessment and it has positive washback effect on everyday teaching (Hamp-Lyons, 1992, p. 9).

Variety of writing assessment instruments has been supplemented by computer-based assessment recently, which is still under construction, as there are several features of the testing process that cannot be modelled with computer programmes. There are two distinct applications of computers in testing written performance. Firstly, a distinction is made between hand-written and word-processed scripts, as using word processors in written performance assessment can result in differences in scores. Secondly, reading scripts and assessing them with a computer programme seems to be close in time, but so far development of a computer programme for rating written products has not been completed. Technological advance influences written performance assessment, however regardless of which aspect is considered, application of computers requires certain level of computing literacy, and adequate financial background (Weigle, 2002, p. 237). Computers are essential technical tools and as they are used in almost all walks of life, writing on a computer has become natural. However, written performance assessment is still mostly accomplished with paper-and-pencil tests. Lee (2004) conducted a study in which he compared holistic scoring of paper-and-pencil and word processed texts. The rationale for using word processors in written performance assessment is related to the way writers produce texts in real-life on computers, which seems to be closer to process-oriented writing. Lee summarises research related to using computers in L2 language assessment so far and concludes that sometimes results are contradictory that is why further investigation in the area is justified. One of the research questions in Lee's study is related to raters' reactions to differences between handwritten and word processed scripts. The issue of raters' reaction to layout features is addressed. Lee's findings show that raters attend to layout features and their judgements are influenced by the appearance of the script. Lee found that inter-rater reliability is much lower in rating handwritten scripts than word processed ones (Lee, 2004, p. 11). It shows that raters cannot distance themselves from text appearance. Shaw (2003; 2004) draws attention to the fact that changing the medium in a test can affect the final score and it raises construct validity questions. He gives an account of a study in which handwritten scripts were typed and, then both typed and the original handwritten ones were scored and the scores analysed. Findings of the study show that although raters found the typed scripts easier to read, the typed versions seem to deflate the mean score. Shaw attributes this to the characteristics of L2 written performance assessment in which there is more attention paid to linguistic features of a text than to its content (Shaw, 2003, p. 10).

Computer software programmes for rating written performance exist and have been used by some high-stakes testing institutions. Rudner and Gagne (2001) compare three scoring programmes used mainly in the United States, where the first software was introduced in 1966. They are different in precision of written performance assessment, but as they mostly focus on surface features, such as word counts or frequency computing, content features are not assessed properly. However, compared to human rating it is consistent and quick. Therefore, considering advantages and disadvantages, Rudner and Gagne suggest that automated and human scoring should be used together to save costs and energy to arrive at reliable scores. Another computer scoring system called Latent Semantic Analysis (LSA) is based on semantic text analysis and is a reliable measurement tool for content subjects. E-rater is designed similarly to human rater behaviour and is used operationally to assess written performance. Although computer scoring has undergone considerable improvement, for example, using E-rater programme in a high-stakes assessment context in the US, they cannot substitute human raters entirely (Hamp-Lyons, 2007; Weigle, 2002, pp. 234-237). A detailed discussion of the issues would be beyond the scope of the focus of the study presented in this book.

Self-assessment is not an innovation in language testing, students can assess their language ability either with objective tests or rating scales with the so-called "can do" statements. These assessment types measure language ability receptively; language production cannot be measured this way. Brown (2005) proposes a model for the self-assessment of writing performance. She presents a study in which students assessed their own writing skills with the help of annotated samples. The model originates in rater-training during which raters standardise and practise rating on samples annotated on the basis of benchmarks. Brown argues that students' writing improved considerably when they compared annotated scripts and self-assessed their own writing.

3.3 Nature of Written Performance Assessment

3.3.1 Validity of Writing Assessment

Research agenda in written performance assessment in both L1 and L2 focuses on different aspects of validity that comprise issues related to the task, the writer, the scoring procedure and the rater (Hamp-Lyons, 1990, pp. 70-73). In order to be able to make valid inferences on candidates' performance, sources of variability should be investigated and irrelevant variance should be excluded as much as possible (O'Sullivan, 2002, p. 14). Among others, construct validity seems to be the most important, which should be established on theoretical and

empirical basis. Face validity, or task resemblance to real language use, content validity and several aspects of criterion validity should also be considered. In what follows, I will discuss the most significant characteristics.

As mentioned in Chapter Two, the key question in establishing validity of a test is defining the language ability or the construct that the test is intended to measure. Test quality can be maintained only if the test is valid and validity should be examined continuously from several aspects (Shaw & Jordan, 2002, p. 10). Alderson (2000, p. 118) says that it is not easy to define construct as it is an abstract entity. The construct can be under-represented if there is a limited number of elements present in the construct. The other factor that should be considered is construct-irrelevant variance of the test, which is a kind of measurement error caused by factors not relevant to the construct itself, such as scoring errors or background knowledge required (Weir & Shaw, 2006).

In examining validity of construct a distinction can be made between convergent and divergent or discriminant validity. The former refers to inferences based on measures that give similar results, whereas the latter relates to those that give different results. Three constructs of written performance were examined to reveal details of validity of inferences based on them (Spencer & Fitzgerald, 1993). The three constructs are structure, coherence and quality, which were examined with reader-based measures, i.e. from the reader's perspective, and with text-based measures, i.e. examining the text features. Spencer and Fitzgerald (1993, p. 225) explain that they found clear convergent validity for structure, but validity for coherence and quality are not evident. They conclude that the construct should be defined clearly not only for the task, but for the scorers as well, and it is especially true for quality and coherence measures. Raters have to have a straightforward understanding of the construct in order to be able to assign scores on the basis of which inferences can be made. They call for more research in the area of readers' understanding. This issue of raters' script interpretation is addressed in the present study.

Assessment of written performance for a long time was mainly product oriented in which writing appeared to be decontextualised and thus distant from real-life context. Recognition of writing as a social act has brought the notion of specifiable context to the surface, meaning that if writing is considered to be a social act, then socio-cognitive factors should also be considered. It means that the writer should be aware of the audience, the purpose of writing, and the task requirements. Thus, test design should take the test-taker, the purpose and the real-life situation into account (Weir & Shaw, 2006). Weir and Shaw call this relationship "symbiotic" and say that "the 'symbiotic' relationship between context validity, cognitive validity and scoring validity constitutes what we refer to as *construct validity*" (emphasis in original, 2006, p. 11). They examine five Cambridge tests to find out to what degree tests can ensure the identification of different levels of proficiency and what evidence of the underlying language

ability is present. The conclusion is that tests should be explicit in their construct definition to make the results comprehensible for the public.

As mentioned above, writing performance assessment involves actual writing and is different from traditional testing methods, in which the test taker's language ability is observed using a measurement instrument on the basis of which a score is assigned and inferences are made about the ability in question. This test type is called "timed impromptu writing test" by Weigle (2002, p. 59), who suggests that the term "direct" is problematic for this kind of writing test and she highlights the fact that the test is taken within a prescribed time limit on the basis of a task or tasks unknown to the test takers. This feature of written performance assessment raises concerns regarding validity and time factor is a feature that should be included in providing evidence for the validity of written performance test.

Factors that characterise performance assessment in general and writing performance in particular include the measurement instrument, the rater, the rating scale, the performance and the candidate (McNamara, 1996). It follows that issues related to construct validity of written performance include the task which elicits writing performance; the scoring method with rating criteria as well as readers' understanding of these criteria (Weigle, 2002, p. 51). These factors interact with each other in a complex way and they are considered one by one in the following sections.

3.3.2 Task Characteristics

Similarly to real-life tasks, tasks for writing performance assessment are various depending on the context they appear in. Although a testing situation is different from real life situations, tasks should elicit a representative sample of the language (Spaan, 1993). Tasks should be constructed without ambiguity and they should elicit the language consistently through several administrations. The subject matter or the content of the task is what the test taker is expected to write about; in content-based tests the language domain to choose from is narrower. In general tests the content should be familiar to the test takers so as not to influence performance. Language is generated by a stimulus or a prompt, which can be either text or non-text.

Writing tasks can be characterised from several aspects: focus of measurement of a writing task can be indirect or direct. An indirect task measures the language ability indirectly, focusing on discrete elements of writing ability, such as editing, grammatical structures or cohesive devices. Direct tasks are intended to measure language ability and reflect the communicative feature of the ability; direct writing test tasks require the candidate to produce a piece of writing as in a real communicative situation. As language performance assessment focuses on observing candidate behaviour on tasks which serve as a basis for making

inferences about language ability, direct tasks are more suitable for writing performance assessment.

The measurement instrument or the test task for writing performance assessment should sample the construct as closely as possible. This notion is called face validity and it can be easily maintained using direct tasks that resemble real-life writing tasks (Hamp-Lyons, 1990. pp. 70-71). It means that writing tasks have to be authentic. In written performance assessment there is either a limited amount of tasks or a choice is offered; therefore, special attention should be devoted to task characteristics. In order to elicit sufficient language for generalisability, assigning more tasks is possible. In this case a choice is offered and the tasks should be similar in difficulty to measure the construct in a similar way on different administrations (Weigle, 1999, pp. 145-146). If there are more tasks assigned a further question is whether to use the same rating categories or design task specific ones (Hawkey & Barker, 2004).

The shift in theoretical approach to written language production affects the way writing ability can be tested. Growing interest in the ways written language is produced resulted in introduction of the so-called process-oriented approach in language instruction. Process approach, as discussed in Chapter One, centres on the way language is produced, so it resembles writing in real-life situations. However, product-oriented approach to written language production lays an emphasis on the content of the text, not on the processes writers go through when composing. It is especially important in academic contexts where assessment focuses on the content (Weir, 1993). The importance of the writing ability in academic context is highlighted by research into possible task features in a degree programme (Taylor, 1996). Summary of task characteristics consists of five areas, each of which is described in detail. The areas are the following: place of the test administration; length of the product ranging from half a page to an extended piece above 10 pages; genres, such as an essay, a report or a case study; cognitive demands and rhetorical specifications (Taylor, 1996, p. 122). Considering process- and product-oriented approaches in written language production seems relevant in testing writing ability. Process approach characteristics include writers' collaboration and unlimited time. These two features raise reliability concerns in testing: if there are more writers it is not clear whose ability is tested. Similarly, there can be problems with time limits: if there is no time limit reliability is affected (Weir, 1993, pp. 133-136).

Still, the cognition-based view of writing should be reflected in tasks assigned to testing writing ability. Good practice in language testing in general and writing in particular involves considering three aspects: conditions, operations and quality of output. As far as tasks are concerned, Weir provides explanations illustrated with examples how the three aspects can be realised when eliciting written performance. He proposes a framework for testing written performance in which conditions and operations are the situational and instructional features

compiled into a checklist for test constructors and evaluators. The quality of output constitutes the assessment criteria. He also says that students should produce extended texts of at least 100 words long, however assigning more tasks increases reliability, so more than one task is advisable but students should write under timed circumstances (Weir, 1993, pp. 133-136).

Cumming, Grant, Mulcahy-Ernt and Powers (2004) conducted research into how instructors perceive prototype TOEFL tasks. Instructors evaluated their students' performance on prototype tasks and compared them to their perception of students' knowledge. Results show that the content of high-stakes test tasks corresponds to instructors' perception of the construct in question. There are some issues raised in connection with task types. Integrated tasks have advantages and disadvantages: they serve authenticity well, as they represent language use as it occurs in real-life situations. However, if students miss the content of the input, the reading or the listening material, they are not able to complete the writing or the speaking tasks successfully. This happens especially with low achievers.

Task dimensions include components, such as genre, rhetorical tasks and cognitive demands. In addition, the length of the text required and the time allowed together with the way the prompt is worded are also important. In some cases the writing test comprises one or more tasks, and in others test takers can choose from prompts. Transcription mode also plays a role, as there is a difference between typed or handwritten texts. The prompt can include scoring criteria to inform the test-taker of the requirements. These task variables overlap in some cases and they cannot be isolated. They show that a variety of tasks that can be assigned in both indirect and direct writing assessment and as these features are present in different proportions depending on the task and the construct. In direct writing tests the number of tasks to be assigned is an issue from different aspects. One task can represent a certain area of language ability; however, if there are more tasks, a decision should be made whether they elicit the same language or different areas of language ability. If there is a choice offered to candidates, each task should have the same features to elicit the same language (Weigle, 2002, pp. 60-62).

It follows that prompts have a considerable effect on written language production. Spaan (1993) gives an account of her findings as far as prompt differences are concerned. The four writing tasks used in her study aimed to provide evidence on how prompts influence writing performance. They differed in rhetorical mode and content and were analysed from four aspects: cognitive demand, writer's intention or purpose, role of the writer and the audience and demands of the content. Comparing the scores of the holistic scores did not show big differences, however, some implications for writing performance assessment are important: if there is a choice provided, candidates choose both the safe and the challenging topic, and the choice is not level specific. Task content plays a

role as well, if the candidate does not have sufficient knowledge of the topic or cannot say much about it, the score is lower.

The effect of task difficulty is examined by Jones and Shaw (2003) in a high-stakes assessment context to find out more about the way task difficulty affects performance. Raters were assigned to assess the same scripts employing different level rating scales. The conclusion is that tasks should be chosen bearing the rating scale in mind, as there is no point in giving difficult tasks at lower levels and easy ones at higher levels. Raters have the knowledge of the construct at their disposal according to the rating scale and they adjust rating not to the difficulty of the task but to the rating scale.

3.3.3 Definition of Audience

The three factors that are considered to play the main role in written performance assessment are the task, the candidate and the rater (Shaw, 2007, p. 17). It is evident that in a testing situation the rater corresponds the audience; however, in order to maintain task validity and authenticity, the audience needs specification. It follows that the prompt of a writing task should include specification of audience, the writer's role and the style required. Definition of the audience is important, as the mental representation of the target audience influences the writer's purpose of composing and the strategies employed.

Wong (2005) followed four writers' composing processes on a task in which the audience was not identified. He found that writers developed different images of the audience, two of them wrote bearing in mind their lecturer as an audience, but one perceived the lecturer as an evaluator, while the other as a feedback provider, one writer wrote for a less knowledgeable audience, and the fourth for herself to reflect on the issue. Wong concludes that the audience perception defines writers' purposes: they either write to display knowledge or to facilitate learning to write and they adjust strategy use according to the purpose they have in mind. Another study provides empirical evidence for the role audience's age plays in written performance. Porter and Sullivan (1999) give an account of a study in which they examine the content and the layout of letters written to people different in age. They conclude that writers' awareness of the audience influences performance; the more respected the perceived audience, the more detailed, sophisticated and careful the writing appears to be. Although the context of the study is Japan, where age is a significant cultural factor, findings highlight the need for careful consideration of the designated audience in written performance tasks.

3.3.4 Test Taker Characteristics

Although test takers play an important role in the assessment process, little is known about their approach to the writing task and their attitude towards accomplishing it. In writing performance assessment the task prompt should be designed in a way that it elicits the required amount of language from the test taker. Theoretical discussion of the composing process in Chapter One suggests that writing is a cognitively demanding task and this chapter focuses on attempts to reveal how this complex mental activity can be measured.

One characteristic feature of the writing process is the writer's continuous editing of text, which is accomplished according to some internal standards. These internal standards are the writer's internal criteria about the task in progress and the idea of what good writing is (Johnson, Linton, & Madigan, 1994). On the other hand, when a reader assesses the same text, his or her internal standards of good writing come to the foreground and the assessment is influenced by these internal criteria. Johnson, Linton and Madigan asked writers and experienced writing instructors to sort texts independently from each other according to quality, justify their choice and the results were compared. Authors conclude that if the composition instructor's criteria are similar to the writer's criteria, the writer produces a good piece of writing, which means that they share the internal standards for a good piece of writing (1994, p. 242).

Test takers' variables are not defined and examined sufficiently either in L1 or L2 writing ability assessment. Since writing is a personal enterprise, individual characteristics play a role in assessment (Green, 2004). The way language performance is elicited determines test takers' response; they may interpret the task and expectations differently, as they have different language abilities and cognitive styles. These and affective factors together with social characteristics influence test takers' performance (Hamp-Lyons, 1990, pp. 76-78). Weigle makes the same observation as far as research on test taker characteristics is concerned and emphasises the need to look into test takers' thinking as they may misinterpret task requirements and thus, present different quality of language ability (2002, pp. 74-75). It is clear from the definition of language ability in Chapter One that it is not only language knowledge that takes part in task performance, but there are metacognitive factors that interact when producing language, test taker characteristics, especially in written performance assessment where there is a limited number of tasks, should be taken into account in test design.

3.4 Conclusion

In written performance assessment, similarly to the assessment of other skills, a test-taker first faces a task which elicits certain language performance. Then, with the help of cognitive strategies language is produced. Written performance assessment should focus on written language production, as understanding composing processes promotes the development of appropriate measurement instrument. The language sample produced during assessment is a complex phenomenon and varies from person to person. In order to make inferences on people's written language ability we have to construct the task in a way that it elicits a language sample that serves as a base for measurement. The measurement instrument should reflect the construct of writing ability we want to measure so that the raters can judge the performance appropriately. Thus, the score we arrive at depends on several factors and results from the interaction between the test taker, the performance, the rating scale, the rater and the rating procedure (McNamara, 1996).

This chapter overviewed briefly the main steps in the history of assessing writing ability and then, introduced the esential features of L1 and L2 writing ability assessment. As they share similar features, the distinction between the two was made only where it is relevant. Before going into details of written performance assessment I presented some alternative ways of writing ability assessment. Most of them have been developed recently on the grounds of advance in theory of written language production. They mostly complement other assessment methods, as their characteristics raise reliability and validity concerns. Discussion of the nature of written performance assessment was organised within the frame of validity issues and the elements that interact in the assessment of writing were presented one by one.

Finally, I outlined test-takers' characteristics to demonstrate that written language production depends on individuals as well. The nature of written performance assessment includes issues of rating procedures and rater characteristics. They represent the focus of the present study and I will devote a whole chapter to detailed discussion of rating procedures in what follows.

Chapter 4

Rating Written Performance

Introduction

The aim of the present chapter is to further develop the discussion of written performance assessment in Chapter Three and focus on the nature of rating written performance. First, I present various rating scales which aid rating decisions. Raters play a decisive role in rating processes; therefore, their characteristics have to be considered in any rating exercise. Next, I introduce some frameworks of rating processes: one approach views rating as a problem-solving activity similarly to psychology, while the other considers rating as a collection of different strategies. The complexity of rating processes is demonstrated in the discussion of five models known up to date which are presented one by one in the following sections. The last parts of the chapter deal with rater stability and rater training. The view of the role of rater training has changed recently, earlier it was considered to be the only guarantee for maintaining standard in scoring, but nowadays the ultimate trust in it has faded (Alderson, Clapham, & Wall, 1995; Weigle, 2002). Rater training is important to arrive at reliable scores, but apart from this function, it serves standardisation and should have a monitoring role as well (Alderson & Banerjee, 2002).

4.1 Scoring Procedures

To define scoring procedures during test design three main steps should be followed: first, we have to identify the ability we would like to measure theoretically, second, operationally, which means deciding on the way of elicitation; and third, we have to determine the way observations are quantified (Bachman, 1990). Language ability, like other mental abilities is difficult to measure, the measurement instrument should be developed in a way it reflects the construct in question and raters can use it consistently. That is why training raters prior to a rating exercise seems to be crucial, although some researchers, as mentioned above, question its necessity. "No matter how extensive or thorough it [rater training] may be, the rating is still a perception, a subjective estimate of quality" (Purves, 1992, p. 118). It can be a strict statement about the value of rater training, but it shows the complexity of rating processes. It follows that interaction between the rating scale, the rater and the rating process determines

the score, which is then interpreted to describe the language ability in question. In the following section rating scales are examined first in detail to establish the role they play in scoring procedures.

Rating scales serve several purposes and they play an important role in performance assessment, as discussed in Chapter Two. First, they are used to define the level of language ability and second, raters make judgements on performance according to them. The former is called user-oriented or reporting purpose, while the latter is assessor-oriented or rating process guiding purpose. The assessor-oriented purpose is supposed to maintain reliability of rating, ensure standard of rating and provide a common ground for score interpretation by raters. These characteristics are in the focus of attention of the present study. The third purpose is the constructor-oriented purpose, which aids test specification definition (Alderson, 1991b, pp. 72-74). Rating scales are sets of descriptors associated with bands or points to define the proficiency level of language ability. It is possible to make inferences about candidates' language ability with the help of rating scales. Rating scales are used to assess oral and written performance and raters are trained to use the scale in a standard way. In addition, the scales reflect test specifications, so they serve as an aid for test constructors to design test tasks. Rating scales consist of numbers, letters or other labels. Alderson et al. (1995, pp. 107-108) make a distinction between holistic and analytic scales; the former is used to assess performance as a whole, the latter provides the rater an opportunity to look at elements of performance and assess them separately. The three most frequently used scales in written performance assessment are holistic, primary trait and analytic scales. The choice among them depends on whether we want to assign a single score or more scores. In addition, it is also possible to have a specific scale for a particular task, or we can use the same scale for assessing different tasks (Weigle 2002, pp. 109-121).

Holistic scoring is based on overall impression during which raters assign scores comparing the script to a scale with several proficiency levels and each level is described. General impression marking is a variation of holistic scoring during which a single score is awarded without a scale to compare the performance with. Holistic scoring has several advantages: the performance is assessed as a whole; its overall effect on the rater is in the centre. In addition, scoring is faster, as there is only one scale to attend to. The definition for holistic writing assessment provided by Hamp-Lyons (1992, p. 2) comprises five components or characteristics. First, in assessing written performance holistically, candidates should produce a text or more texts of at least 100 words each. Second, although the rubric defines the task and provides some kind of prompt, candidates have considerable freedom in composing texts. Third, rating is realised using one or more raters. Fourth, rating is based on a consensus between raters or on sample scripts or on a rating scale. Fifth, raters' judgements

are expressed in scores, which can be later interpreted to make conclusions on candidates' language ability. It means that there is a piece of writing elicited with a writing task and assessed holistically to arrive at a score. This way of rating does not provide sufficient information on the components of language ability, as a single score cannot reflect all aspects. Therefore, it is not widely used in L2 writing assessment. Moreover, criteria interpretation can cause problems, as raters may attend to different features and arrive at different scores, which are then the only sources for making inferences about the language ability. A more detailed description of test takers' writing ability can be given using primary and multiple trait scales or analytic scales (Weigle, 2002, pp. 112-114).

Primary trait scale is one type of holistic scales along with multiple trait scoring methods that are mostly used in L1 assessment. "The theory is that every type of writing task draws on different elements of the writer's set of skills, and that tasks can be designed to elicit specific skills" (Hamp-Lyons, 1992. p. 8). A truly holistic scale uses a combination of main features of a script and ends up in one score. Multiple trait scoring focuses on different aspects of scripts and makes awarding more scores for each script possible. A primary trait scale focuses on key features of a single script and is task specific (Cumming, Kantor, & Powers, 2002, p. 68). Students are assigned to write, for example, a persuasive essay and are assessed on the degree of fulfilment of the task only. The rubric for each writing task contains the task, the rhetorical feature, expected performance description, a rating scale, samples and explanations. Although this way of writing performance assessment would be beneficial in L2 context, it is not widely used (Weigle, 2002, pp. 110-112).

Analytic scales "require the rater to provide separate ratings for the different components of language ability in the construct definition" (Bachman & Palmer, 1996, p. 211). They give more detailed information on candidates' performance than any type of holistic scales described above. They are divided into several aspects of writing performance, for example, content, vocabulary, accuracy, etc., each of them is assigned descriptors according to proficiency levels. Analytic scales are divided into several proficiency levels as well, which can have same or different weighting depending on the importance of the aspect in question. A candidate may get higher or lower scores on different aspects reflecting the differences in the components of language ability (Alderson et al., 1995, p. 108). It is especially useful for L2 learners to diagnose the areas of language ability that need further improvement. There are scales, which consist of several subscales to make more elaborated assessment possible. The score can be either combined or reported separately. Weigle (2002, pp. 114-121) emphasises the importance of explicitness of scale descriptors and clarity of distinction between levels. As a result, scores reported show an informative picture of test takers' language ability. Weigle finds rater training very important regardless of the scale chosen, as training for using the rating scales ensures reliability of judgements. Cumming

and Riazi (2000) examined indicators of achievement in writing in L2 and they chose an analytic rating scale and they found that it showed the elements of language ability in a more detailed way than any other assessment instrument. They also claim that an analytical rating scheme is multi-faceted similarly to the ability it intends to measure.

The dilemma of which rating scale to use in written performance assessment has been intriguing research for some time now. The choice is narrowed to either analytic or holistic scales, as they seem to be appropriate for measuring written performance, as far as reliability and validity of assessment are concerned (Shaw, 2002c). Comparison of the two scales on six qualities of test usefulness framework is proposed by Weigle (2002, pp. 120-121). They are discussed in Chapter Two, and they can help to decide on the type of rating scale for a particular test. These qualities are reliability, validity, practicality, impact, authenticity and interactiveness (Bachman & Palmer, 1996, pp. 17-43). Weigle applies these qualities of test usefulness to compare analytic and holistic scoring and summarises the main characteristics of these qualities as follows. Reliability of a holistic scale is lower than that of an analytic one. Comparison of construct validity shows that a holistic scale is built on an assumption that writing ability develops evenly, whereas an analytic scale can make a distinction between the elements of ability, which means that it is more informative especially for L2 writing performance assessment, as steps of elements of the ability can be followed more thoroughly. A holistic scale is more practical, as scoring is fast and easy in comparison with time and expenses needed for analytic scoring. Impact issues include score interpretations; holistic scores provide less information on writing ability and make decisions more difficult. Authenticity is higher in case of holistic scoring as reading scripts resembles real-life use more than reading scripts for analytic scoring. Weigle does not compare the two scales on the basis of interactiveness, as interactiveness by definition relates to the relationship between the test and the test taker, so it is not applicable for rating scale selection (Weigle, 2002, p. 121).

Empirical evidence supports the above mentioned differences; for example, Nakamura (2002) reports on a research that aimed at comparing the two scoring methods. Three trained raters assessed ninety scripts both holistically and analytically and the results were compared bearing in mind the theoretical framework of the differences discussed above. Findings shed light on the main issues related to scoring procedures. First, Nakamura says that for economical reasons holistic scoring seems to be less costly. Second, she draws up a rank order of choices for scoring procedure regarding the number of raters, the number of writing tasks and the rating scale type. She concludes that the best practice is to have multiple raters and multiple items, and the least appropriate choice is one rater and an impressionistic scale. The second best practice is to have one evaluation with more raters. This option is elaborated in more detail in the

present study to find out more about scoring processes. Third, Nakamura refers to classroom applications of scales and warns that holistic rating can lead to misinterpretations of student ability. Fourth, maintaining construct and content validity and reliability are the most important concerns in making a choice.

To sum up, the choice between holistic and analytic scoring methods depends on the purpose of the test and financial consequences, not simply on quality differences between the two.

4.2 Rater Variables

Rater characteristics constitute another important factor in scoring procedures of performance assessment. The score in fixed response assessment depends on test takers' choice from the offered options and the criteria for correctness are determined in advance. In performance assessment test takers are assigned a task, which elicits certain behaviour that is judged according to rating criteria. Alderson et al. (1995, p. 128) emphasise the role of raters and say that it is of paramount importance that the score on a test does not depend on who the rater is and how consistent his decisions are. Rater behaviour, as discussed in Chapter Two, is characterised by two types of reliability: one is the consistency among the raters, called inter-rater reliability; the other indicates the consistency of individual raters within themselves, called intra-rater reliability. Raters play a decisive role in the assessment process of written performance, as they get into interaction with the task, the performance and the rating scale. McNamara (1996, pp. 121-122) highlights three sources of variability that influence the final score: the test taker's ability, task characteristics and the rater variability. Rater variables can be attributed to the variables of the rating process itself and to raters' background characteristics (Weigle, 2002, pp. 70-72).

Rater reliability should be established before rating takes place during rater training and it is necessary to monitor it throughout rating. As large-scale high-stakes testing involves several raters and a lot of scripts to rate, rating is done centrally or in raters' homes. Maintaining reliability with many scripts can be realised in different ways: there can be samples chosen and remarked by chief examiners or re-marking can be done routinely (Alderson et al., 1995, pp. 129-135). Rater reliability is established in rater training or standardisation meetings as discussed below. Connor-Linton (1995) emphasizes the importance of looking at the rating process from the raters' perspective examining how they make their decisions. Concern about the reliability of rating scales has oppressed the role the raters play in interpreting them. He says that no matter how reliable the rating is if we do not know why raters awarded certain scores, we cannot interpret them. There are three ways of getting an insight into raters' thinking: data can be collected by using think-aloud method, or ethnographic observation, and ratings

and textual characteristics can be compared. He lists some considerations for research design, these are the number of raters and scripts involved, and the amount and the depth of feedback provided. Investigation into rating processes helps to improve rater training, and it also promotes better understanding of the relationship between teaching and testing, which is called washback.

Raters' role in rating processes is emphasised by findings of a study that attempted to examine how raters interpret the assessment scale criteria (Lumley, 2002). The conclusion is that raters play a crucial role in the rating process: they decide on scale interpretation, on solution of occurring problems, and they justify their judgements based on scale descriptors and knowledge gained during rater training (Lumley, 2002, p. 267). It means that the interaction among the elements of written performance assessment is controlled by raters, and their judgements are their interpretations of the scale compared to the performance and their expectations. Lumley explains the need for rater training with focus on reaching a common understanding of what the scale means. He says that rating "does not require special training to read and form some sort of judgement of a text, but rating is considerably more complex than this" (Lumley, 2002, p. 268). This implies the need of investigating further raters' decision-making processes and finding out more about how raters arrive at a score. The present study centres on finding out further details of the interaction and attempts to contribute to existing research.

Interaction between the task, the performance and the rating scale is operated by raters who can interpret the task differently; they have their own expectations as far as the task completion is concerned. They have different backgrounds, are of different age, have their own perception of the performance: these characteristics influence the judgements they make. This kind of interaction is called "rater – item" interaction. Another type of interaction is the "rater – candidate" interaction, which is the influence of a particular candidate or group of candidates on the rater. One element of the composing process is editing, during which writers evaluate their texts according to own internal standards or criteria. Similarly, raters evaluate texts using their own internal criteria and they compare their criteria to those of the writers. It follows that if the two sets of criteria are close to each other, readers' agreement regarding the quality of written text is higher than in the case of different internal standards (Johnson, Linton, & Madigan, 1994). Although these findings result from research on improving writing, they are relevant to rater behaviour as well.

The rating scale, as discussed above, can also influence raters; they can attend to scoring categories differently, giving more emphasis to some of them and less to others. Rater types can be identified depending on which part of the rating scale they attend to more: some tend to assign scores from the middle range of the scale, while others choose from the two extremes. Interpretation of the scale descriptors can also be a source of discrepancy, no matter how precise

the wording of descriptors is, their interpretation depends on the reader. Raters can also be distinguished according to their experience; however research so far has not made clear-cut distinctions between experienced and inexperienced raters. McNamara summarises his perception of raters' saying that "raters display certain characteristics in their participation in the rating process, and these characteristics are a source of potentially considerable variability in the ratings of a performance" (McNamara, 1996, p. 127). He adds that these sources of error can be decreased by implementing certain measures, but cannot be totally eliminated. One way of lessening the effect of rater variables is providing rater training for them.

Written performance assessment is characterised by an interaction between the rater, rating scale, rating processes, performance or script, instrument, and the candidate (McNamara, 1996). After an introduction of these elements in Chapter Two and here, the interplay between them is presented in what follows. The elements of rating processes interact, thus it seems possible to attempt to design a framework of the scoring processes to understand the interplay between the elements.

4.3 Rating as a Problem-Solving Activity

DeRemer (1998) approaches rater behaviour from a psychological point of view and sees the rating process as a problem-solving activity in which the rater aims at a goal, a decision about the text quality. Her focus is on the processes that raters go through while evaluating portfolios. DeRemer says that rating is not a simple matching activity of scale descriptors and the text, but the rater develops an own interpretation of the criteria and the script to make decisions, thus rating is a constructive activity. She conducted a study (DeRemer, 1998) based on think-aloud protocol analysis combined with quantitative data, which were the scores awarded. Attempting to understand how raters elaborate on the rating task she identified three areas raters attend to: impression scoring, text-based evaluation and rubric-based evaluation The three main processes identified during rating are search, or finding an appropriate score for the script, and two recognition operations: a simple and a complex one. As there are different processes applied by raters with different focus, such as search or recognition when assigning scores, score interpretation can have different meanings. Thus, there are three identifiable rater behaviour types: one that focuses on the general impression of the text, the other puts the text in the centre and the third attends to the scale.

She is concerned about the validity of scores as well, saying that it is not always clear whether scores reflect the construct. As for the reliability of scoring, DeRemer (1998, p. 26) says that scoring with traditional scales and

after traditional training does not necessarily end in rater agreement. As writing assessment is not about making a choice between right and wrong answers, it is an ill-structured problem-solving task and it is not possible to create a standardised rating process. Thus, rater training is about reaching agreement among raters on the basis of benchmark scripts and still, raters have to develop their own procedures for rating texts. She says that raters, when employing some kind of a rating scale to evaluate various texts, go through a constructive activity which results in different judgements.

Although research introduced above represents a considerable contribution to investigations into rating processes, it has been criticised as soon as it was published. Torrance (1998) acknowledges the significance of the study, but questions the validity of research methodology and depth of the analyses conducted. The implications of Torrance's points highlight the importance of conducting research into raters' thinking and decision-making processes and he says that educational research is a complex area in itself and should rely on other sciences, such as psychology. He refers to the relevance of this study to education saying that testing can have serious consequences on curriculum; it can narrow it down if not conducted properly. Torrance encourages further research in the area and raises questions in connection with findings on rater thinking discussed in DeRemer's study.

Scoring procedures for traditional and performance assessment differ significantly, as in the former counting is mostly used, while in the latter the observed behaviour is evaluated using judging processes. In performance tests test takers' language performance is elicited, and a rater or raters judge it employing a rating scale (McNamara, 1996). Raters arrive at some kind of decision which can be expressed with a score; then, the scores are interpreted and inferences are made on test takers' language ability. Investigation into the processes raters go through and arrive at a score involving judgements is one of the major concerns of researchers and examination bodies (Shaw, 2001, p. 2). The investigation of raters' decision-making processes is in the centre of the present study and the intention is to shed light on their judging processes.

The dichotomy of counting and judging seems straightforward, as in an objective test there is no need to decide on the quality: an answer is either correct or not (Alderson, 1991a). On the other hand, in tests where language production is required, the quality of the product needs judgement. Pollitt (1991) uses a "sporting" metaphor, and compares rating processes to sporting in a sense that there are observations that can be assessed by counting (e.g., time in races), while others by judging (e.g., figure skating). The same notion is referred to by Jones and Shaw (2003, p. 12) who distinguish between difficulty and quality of written performance. They refer to language proficiency frameworks in which levels are identified in terms of difficulty, i.e. what a candidate can do, and quality, i.e. how well a candidate can perform on a task.

There are several models of judging and decision-making in psychology, the one that can be related to rater behaviour is a dual processing model (Greatorex & Suto, 2006). There are two types of cognitive operations: quick and slow thought processes. The former are used when comparing the answer to a model answer in the marking scheme, the latter are used when the answer needs consideration as it is different from model answers provided. Greatorex and Suto say that transition from one operation type to another is possible, when, for example rating categories are internalised. They analysed GCSE raters' verbal protocols and identified five cognitive marking strategies: matching, scanning, evaluating, scrutinising and no response (2006, p. 6). Matching strategy entails comparison of performance to the answer on the marking scheme which is either written or memorised; it can be used with short answer item type. Scanning strategy sometimes involves more scans of the text for different details; it can be either pattern recognition or semantic processing or both depending on the text and the success of scanning. This strategy can be used at the end when the score is awarded as a checking strategy. Evaluating strategy involves thought processing and a combination of rater characteristics during which the examiner decides if the response is correct or not. Scrutinising strategy occurs in case the response is unexpected and there is a need to establish what the candidate wanted to do (Greatorex & Suto, 2006, pp. 8-12).

Research into rating strategies is by far not exhaustive; there are several questions that need to be investigated at more depth, for example, approaches raters use in marking, raters' focus on different text elements, inter-rater consistency and how the level of the text influences rating (Shaw, 2001, p. 2). These issues have generated further inquiries, such as building frameworks of scoring processes.

4.4 Frameworks of Scoring Processes

Scoring or rating involves cognitive processes that depend on several factors and as is the case with other mental abilities, they are not easy to trace. Research that has been conducted into mental processes raters go through when assessing written performance focuses on different aspects of the processes and outcomes and makes inferences on both qualitative and quantitative data. Decisions raters make are influenced by personal characteristics such as age, experience or professional background. Thus, differences in raters' judgements stem from three possible sources: the raters' understanding and interpretation of scoring criteria, their interpretation of the text they assess, and the differences in rating processes (Wolfe, Kao, & Ranney, 1998, p. 469).

Apart from these features, the investigation of rating processes comprises the issue of sequencing, whether raters go through a linear process, or it is recursive

or both depending on different features. In getting a deeper insight into these processes, qualitative data are collected using some kind of introspection, while quantitative data comprise statistical analyses of scores to make conclusions on marker features based on scores awarded. Results of such research seem to show contradictions in some cases to underline complexity of the issue and multiplicity of factors influencing scoring processes. Main lines of research in the area of rater behaviour focus on raters' attention to the text features, such as content, mechanics or organisation, their scale interpretation, raters' perception of task difficulty, their expertise in rating and even their social background (Cohen, 1994b, pp. 332-336). The most recent frameworks are presented in chronological order to exemplify the complexity of factors that appear in rating processes.

4.4.1 Milanovic, Saville and Shuhong's Framework of the Scoring Process

Raters employ several reading approaches when assessing written performance and they focus on different elements of scripts, the extent of this attention may vary. Milanovic, Saville and Shuhong (1996) conducted research to reveal raters' reading approaches to find out what raters pay attention to and to what extent during rating. First, they compiled a model of decision-making processes built on earlier findings; they mostly relied on the model developed by Cumming (1990, cited in Milanovic et al., 1996, p. 94) and examined how four groups of 16 raters rated holistically two different level scripts. The decision-making model attempts to reflect both the rating processes and the elements raters focus on during rating. Thus, as Figure 4.1 shows, the rating process has four stages: pre-marking, scanning, quick reading and rating. Each of the four stages involves different focus and rater behaviour.

During the first, pre-marking stage raters internalise the marking scheme and interpret the task. Then, in the scanning stage they focus on surface elements of scripts, such as length, format, handwriting and organisation. When raters read the scripts quickly, they focus on overall comprehensibility. The rating phase is characterised by focusing on both the content and linguistic features of texts. Content is assessed for text relevance, topic development, coherence and organisation. Linguistic features focus on looking at errors, assessing syntax, lexis and spelling (Milanovic et al., 1996, p. 95).

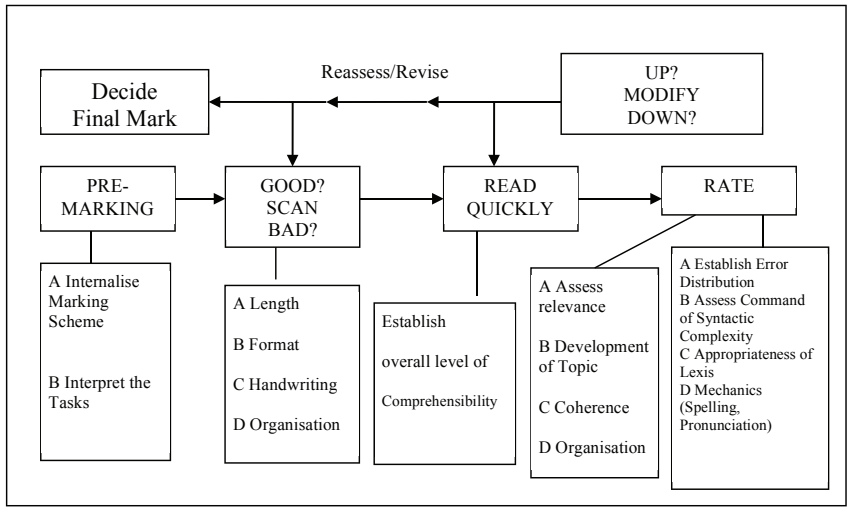


Figure 4.1. A model of the decision-making process in composition marking (Milanovic et al., 1996, p. 95)

Milanovic et al.'s findings reveal that raters use four different reading approaches in different proportion. One is a so-called principled read, which means the script is read twice bearing in mind some of the scoring criteria; whereas during reading for the second time attention is paid to deciding the score. The next approach is the pragmatic read, which also involves two readings, but in this case the second is motivated by solving some encountered difficulties in scoring. The read through approach is not detailed and it is employed especially with short scripts. The fourth, provisional approach identified comprises of a quick read after which a provisional mark is awarded.

The attention raters pay to the content and linguistic elements of scripts is very varied and findings show that raters do not reflect similarly on what they read. They mention that they focus on what makes scripts different from others or remark on positive or negative features only. Milanovic et al. (1996) conclude that raters express the highest degree of subjectivity when assessing content. The differences can be attributed to several factors, such as the level of the scripts, and raters' focus on vocabulary and content more with higher level scripts and they attend to effectiveness and task completion with intermediate level scripts. Moreover, raters' background seems to play a role, the four raters had different background: markers of two levels of an EFL examination, EFL teachers and mother-tongue markers. Intermediate examination markers focused more on length, while higher level markers on content and markers of L1 writing focused on tone of the scripts. Milanovic et al. acknowledge that these general

conclusions need further research into rater variables and propose a design for further research (1996, p. 106-107).

4.4.2 Wolfe's Framework of the Scoring Process

In order to make inferences about rater behaviour several research methods are used. Wolfe (1997) attempts to make inferences about raters' reading style and proficiency, as he calls inter-rater agreement. He summarises research conducted earlier and concludes that earlier studies focused on the content of assessment not the procedure and recently attention has turned to examining what goes on in raters' mind. He refers to the information-processing model set up by Freedman and Calfee (1983, cited in Wolfe, 1997, p. 88) in which there are three identifiable processes: the rater first builds an image of the text; then, this image is assessed; finally, the rating is articulated. It means that rating depends on the rater's interpretation of the text, which can vary from rater to rater (Wolfe, 1997). From the beginning of the 1990s research into raters' thinking identified the features of scoring processes that are influenced by raters' background, their expertise in rating and the training they go through.

As in most preceding research, Wolfe tries to identify the differences and similarities between raters of different rating "proficiency". He makes a distinction between proficient, intermediate and non-proficient raters on the basis of rating agreement, thus the emphasis is on raters' ability to come to an agreement. He builds a model of rating process on the basis of this distinction and proposes a refined model of scorer thinking considering earlier research in the area. The model consists of a framework of writing and scoring (see Figure 4.2), the former focuses on processing actions, the latter on content.

First, raters interpret the text and create their own image of the text which is evaluated and the decision justified. During the scoring process raters focus on the content features of the text to a different degree bearing in mind criteria of the rating instrument. Processing actions comprise of text interpretation involving reading and making comments on text; evaluation constitutes monitoring, reviewing and making decisions; justification actions are diagnosing, coming up with rationale and comparing texts. Content focus actions involve comments on appearance, assignment, mechanics, organisation, story telling, and there are non-specific comments as well. The model suggests that raters make up their own image of the text which they assess according to their interpretation of the scoring rubrics.

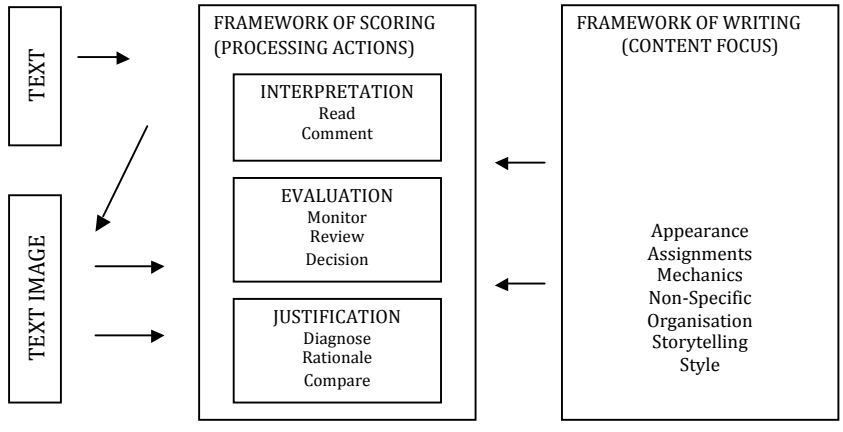


Figure 4.2. Model of scorer cognition (Wolfe, 1997, p. 89)

Wolfe investigates three hypotheses that he formulated based on the model of scoring. First, he investigates if more proficient raters make fewer jumps between content focus categories than less proficient raters. Second, he investigates the differences between the numbers of early decisions made by different raters. Third, he is interested in personal comments. The research is conducted using think-aloud verbal protocol analysis. While examining the jumps raters make Wolfe concludes that less proficient raters make more jumps which indicate that their decision-making processes are more chaotic.

As far as early decisions are concerned, the protocol analysis reveals that less proficient raters tend to make decisions during reading and not at the end, as proficient raters do. Proficient raters make fewer personal comments, which fact in Wolfe's interpretation means that as rating is a cognitively demanding task, less proficient raters find it difficult to cope with the task and thus they often deviate from the rating process. They tend to focus on surface features or break the evaluation down into chunks, which run contrary to the marking scheme, which is holistic marking in his study. Findings of the study can be utilised in rater selection and training, and further research into raters' thinking processes is justified (Wolfe, 1997).

4.4.3 Lumley's Framework of the Scoring Process

The model of the rating processes proposed by Lumley (2000; 2002) is based on findings of a study he conducted in a large-scale testing context with experienced raters. Four raters had to asses 24 scripts using an analytic scale

and produce think-aloud protocols. A coding scheme was first developed on the basis of the collected data which was followed by data analysis to establish a model of rating processes. Central to data processing was to identify and group rater comments, which resulted in three broad categories: management, reading and rating behaviours (Lumley, 2000, p. 134; 2002, p. 254).

The model of rating consists of three distinct stages (see Figure 4.3): the first is the so-called pre-scoring stage, during which raters attempt to get an overall impression of the text focusing on global and local features without identifying scores.

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 4.3. Model of the stages in the rating sequence (Lumley, 2002, p. 255)

The first stage comprises technical comments on script identification and attention is paid to surface features, such as layout and handwriting. The second stage involves raters' consideration of scale categories and they focus on both the text and the scale descriptors one by one to award a score. It is the actual rating, where raters make their judgements and justify them, refer to descriptors, provide examples, and reread the text if needed. The third stage involves score consideration, revision or confirmation and is characterised by finalisation of scores, in which confirmation and revision can be involved. Raters may proceed during the rating process in a linear way or cyclically, they go back and forward between the stages shifting focus between scoring criteria and text parts (Lumley, 2002).

Regardless of sequencing of stages, raters pay equal attention to the four scoring categories and they examine each of them thoroughly. As far as rating criteria are concerned, Lumley (2000, p. 196) confirms what has already been stated by Wolfe (1997) and DeRemer (1998): raters attempt to interpret the scale descriptors to match their impression of the text. Although Lumley finds a close match between texts and scale descriptors, he admits that we still do not know enough of the reasons why raters choose particular scores. However, as

there is no one to one match between scripts and scale definitions, raters often face problems in assessment.

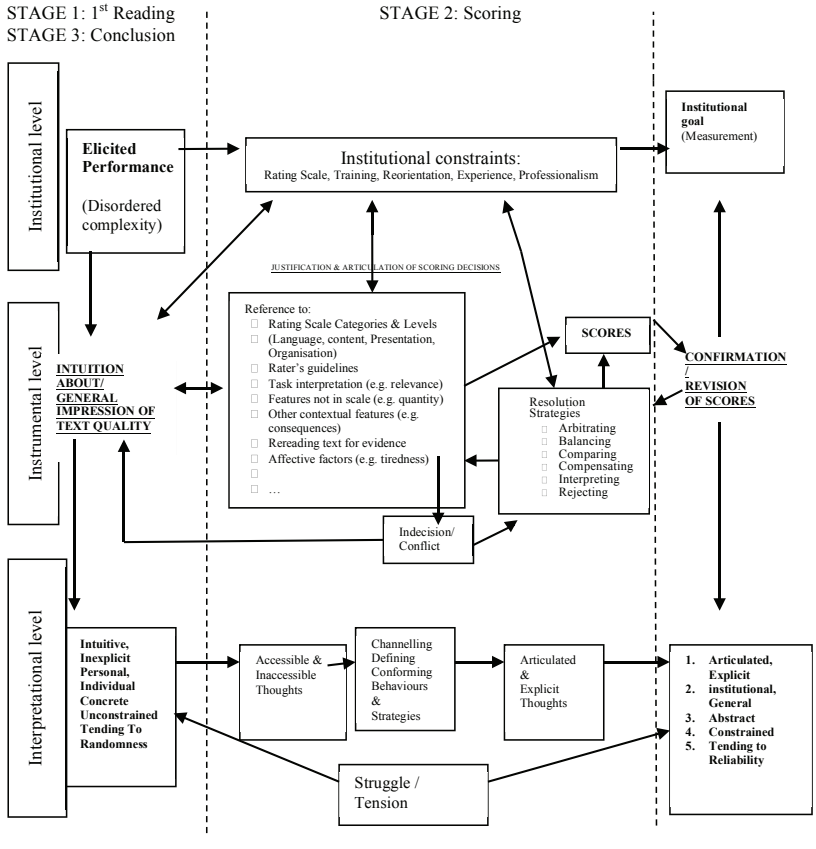


Figure 4.4. A model of the rating process (Lumley, 2000, p. 289)

The model of rating stages discussed above is a simplified version of a more detailed model of the rating processes which includes further features as well. Apart from the three stages of first reading, rating and concluding, there is a distinction between three levels according to which rating behaviours can be identified. The three levels as they appear in Figure 4.4, are institutional level, instrumental level and interpretation level. The instrumental level at the centre of the framework is what research has focused on so far: raters read the text for getting a general impression. The scoring stage of this level is characterised by raters' attending to the rating scale, guidelines, task relevance, other features not included in the scale and rereading the text if they find it necessary. Then,

they arrive at a score, however in case of indecision or conflict raters turn to resolution strategies. The conclusion stage involves score confirmation or revision. Institutional level of the framework is a kind of a social filter, as Lumley calls it and it includes constraints issued in a certain context in which rating takes place. The elicited written performance before rating constitutes a disordered complexity without any particular meaningful quality reference. The scoring stage of this level comprises elements of the rating procedure related to the context they appear in. The rating scale is designed by experts in a certain context; rater training and reorientation also take place in particular context. Similarly, raters' experience and professionalism have their social relevance. At the end of the process there is a score or a set of scores which are indicators of measurement. The third level, the so-called interpretational level includes behaviours, which are characterised by interpreting the relatively simple scale to the much more complex text and compensating for the mismatches between the two. The level is characterised by processes raters go through in case of tension between the text features and rating (Lumley, 2000, pp. 297-298).

When examining the way raters approach problems during rating, Lumley observes three types of strategies used by raters in case a problem arises (2000, p. 239). Defining strategies are related to the scope of and conflicts within scoring categories, and to the boundaries between levels. Raters in some cases expand the scope of the scale; at other times they relate a feature to another descriptor which means that a rating scale does not always fit the script. Comparison strategies are also used; raters compare scripts to each other or compare one task to another. The third strategy type is when raters try to express their overall impression, they refer to these strategies towards the end of scoring and try to maintain fairness in scoring and consider patterns of scores. As Lumley concludes, "the raters' overall impression, based on their professional judgement, is the primary influence in their assessments, rather than attention to the scoring categories as discrete entities, and that raters employ various strategies to ensure that their scores reflect this impression" (2000, p. 281). This framework proposes so far the deepest insight into rating processes. It sheds light on several features of the rating processes and on the complexity of rater behaviour.

4.4.4 Cumming, Kantor and Powers' Framework of the Scoring Process

Research into raters' decision-making processes has revealed more and more about the rating processes and the way scores are awarded. The point Cumming, Kantor and Powers make referring to the need in developing a comprehensive framework for scoring is that "most holistic and other analytic scales lack firm empirical substantiation in respect to evidence about L2 learners' writing abilities" (Cumming et al., 2002, p. 68). They have compiled their framework

based on three consecutive studies as follows: first, they focused on developing an initial framework for rating behaviours based on raters' decision-making while rating EFL writing tasks. Then, they applied the framework to a similar rating task, but with mother-tongue English raters, and finally, the framework was finalised with new writing tasks to trace the areas not covered earlier. The three consecutive studies used think-aloud verbal reports as main methodology for collecting data. Think aloud protocols revealed two types of strategies employed by raters which finding is consistent with other researchers' results, e.g., Lumley's (2000; 2002).

Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation strategies		
• Read or interpret prompt or task input or both	• Discern rhetorical structure	• Classify errors into types
• Read or reread composition	• Summarize ideas or propositions	• Interpret or edit ambiguous or unclear phrases
• Envision personal situation of the writer	• Scan whole composition or observe layout	
Judgement strategies		
• Decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgements collectively	• Assess reasoning, logic, or topic development	• Assess quantity of total written production
• Consider own personal response or biases	• Assess task completion or relevance	• Assess comprehensibility and fluency
• Define or revise own criteria	• Assess coherence and identify redundancies	• Consider frequency and gravity of errors
• Articulate general impression	• Assess interest, originality, or creativity	• Consider lexis
• Articulate or revise scoring decision	• Assess text organisation, style, register, discourse functions, or genre	• Consider syntax or morphology
	• Consider use and understanding of source material	• Consider spelling or punctuation
	• Rate ideas or rhetoric	• Rate language overall

Figure 4.5. Descriptive framework of decision-making behaviours while rating TOEFL writing tasks (Cumming et al. 2002, p. 88)

The final version of the framework distinguishes between interpretation and judgment strategies, as Figure 4.5 shows, according to which raters' behaviour is categorised as having self-monitoring focus, rhetorical and ideational focus, and language focus (Cumming et al., 2002, p. 88).

Interpretation strategies involve reading texts and comprehending their content. Raters used interpretation strategies with self-monitoring focus in reading and making sense of the input, the task and the script, and speculated on writers' situation. As far as rhetorical and ideational focus is concerned, raters looked at the rhetorical structure, summarised ideas and scanned layout. Interpretation strategies with language focus consider language during which raters classified errors and tried to comprehend unclear text. Judgement strategies are those, which need some kind of evaluation: those with self-monitoring focus relate to macrostrategies for reading and rating, such as deciding on sequence, comparing with other scripts, summarising, distinguishing or tallying judgments. In addition, raters applying judgement strategies with self-monitoring focus consider personal views, devise own criteria, react to general impression, and articulate or revise decision. Judgement strategies with rhetorical and ideational focus are assessing logical structure, task completion, coherence, originality, different aspects of text features, source use, and rating ideas. When dealing with language, interpretation strategies comprise error classification and interpretation of problematic text parts. Judgement strategies with language focus are the following: assessing quality and comprehensibility of the text, error features, vocabulary, syntax, spelling, punctuation and language overall. Although the number of judgement strategies exceeds the number of interpretation strategies, raters used both in a balanced way.

Findings of the three consecutive studies show that raters paid equal attention to rhetoric and language features of scripts. In addition, different weighing may be necessary for rhetoric and language, emphasising rhetoric features at higher levels and language at lower levels of the scale (Cumming et al., 2002, p. 89). These studies provide empirical evidence for several issues raised in connection with rating written performance. There are clearly two types of strategies used: interpretation and judgement strategies. All raters attend to similar aspects of scripts, though higher level texts get more attention on rhetorical features than lower level scripts. In addition, English as mother tongue raters attended more to ideas than to the language features of texts. On the basis of the questionnaire that supplemented the verbal protocols, Cumming et al. (2002, p. 89) conclude that experienced raters with common professional and cultural background share similar rating behaviour.

Frameworks of rater behaviour are compiled based on different studies in a variety of contexts in a varying degree of detail. Still, there are apparent similarities between them which support the idea that rating processes have

observable features and tracing raters' behaviour can reveal more about their decision-making processes. Raters' decision-making behaviour is a complex cognitive process during which they evaluate the text that they interpret before employing rating criteria, which are also interpreted by them. It follows that raters develop their own perception of the text and the rating criteria. Complexity of rating processes appears in rating sequences: raters go through stages of rating differently, there is no standard order of rater behaviour, and there is only one common feature: the processes are recursive in nature.

4.5 Rater Stability

The above described frameworks of scoring procedures include what raters do and how they arrive at a score when rating written performance. The complexity of rating processes and the diversity of rater behaviour justify identifications of rater variety. The sources of rater variability can be described from different aspects. As mentioned earlier, raters' agreement can be considered as agreement among raters (inter-rater reliability), or agreement within one rater's decisions (intra-rater reliability). Consensus agreement shows to what degree raters agree, which is usually examined after rater training. There can be percent exact agreement, which means same scores, or percent adjacent agreement, in which there is one score difference (Alderson et al., 1995). Rater consistency can also be calculated, as some raters can be consistently harsh or lenient in their decisions, and some can be inconsistent in their decisions. The rating pattern is acceptable if raters are consistent in their judgements. Measurement estimates take task, rater, error, and interaction components into account to tell to what degree the score represents the construct (Brown, Glaswell, & Harland, 2004).

Agreement among raters is not easy to maintain, especially if there are several raters involved, as in a large-scale test. Those raters who are inconsistent with their judgements can receive extra counselling or can be excluded from the rating exercise. Scores awarded by raters whose judgements are consistently severe or lenient can be adjusted to compensate for the differences. Congdon and McQueen (2000) find this compensation more important than retraining raters, as they say that training can enhance raters' self-consistency but cannot affect rater severity considerably. However, they argue for continuous monitoring and training even during the rating procedure. Their study was based on a large-scale assessment programme in which 16 raters were followed over a period of seven working days assessing 8,285 performances. Congdon and McQueen examined rater severity changes each day with the last day devoted to rescoring. They conclude that rater severity changes with time which has implications in large-scale testing and

means that raters need constant monitoring and ongoing training (Congdon & McQueen, 2000, pp. 175-176). These findings show that rater effects should be considered and especially in case of high-stakes tests it is inevitable to deal with it as it can affect final scores significantly. It is also important to choose raters on the basis of multiple scoring exercises and inconsistencies should be revealed before operational rating.

Thus, one of the issues related to training raters is to decrease the differences between raters, to standardise their judgements and to compensate for rater severity or leniency. Brown, Glasswell and Harland (2004) present their findings on the basis of two studies conducted to examine reliability and validity of scoring processes in a writing programme. They compared data from three approaches: consensus agreement, consistency in scores and measurement estimates. They conclude that although raters had little experience in rating large amount of scripts outside their own classrooms, they reached agreement in scoring in relatively short time, which was a two-day training event. Their findings support the idea that rater training can result in agreement among raters, but this is not the only solution to decrease rater variance. The main issues of rater training are presented in the next part.

4.6 Rater Training

Rater training is the procedure that usually precedes each performance test assessment. It aims to standardise the rating process (Taylor, 2004, p. 2), familiarise raters with the scoring procedures and prepare them for dealing with unexpected situations during rating. During the training raters use the same assessment instrument as in the operational rating and rate some previously selected scripts that represent both the typical and problematic performances (McNamara, 1996, pp. 125-127). These scripts can be benchmarked scripts with scores awarded by experienced raters to establish standard of evaluation (Davies et al., 1999, p. 15). Training raters is especially important if the number of scripts is high as, for example, in large-scale national tests where depending on the number of scripts a lot of raters are recruited. Training consists of several stages and should be conducted before each administration of the test regardless raters' experience. Alderson et al. (1995) note that for reliability of rating it is crucial to go through rater training even if raters have substantial experience in rating.

The length of training depends on the size of the assessment programme, on time and financial constraints. Prior to rater training sessions for writing performance assessment, a chief examiner chooses scripts to sample consensus and problem performances. The sessions start with familiarising raters with the assessment procedure consisting of the writing task or tasks

and the rating scale. Then, raters deal with consensus scripts first and rate them individually and discuss the scores to reach agreement. It is also possible to have benchmarked scripts accompanied by justifications for the scores. If benchmarked scripts are not available, raters need to agree in their decisions and rating should go on until consensus is reached. When there is agreement on these scripts the problem samples are dealt with. The aim is to provide the raters with strategies they can turn to if the script deviates from the task requirements or do not confirm scale descriptors (Alderson et al., 1995). Weigle (2002) does not question the importance of rater training, she also argues for its relevance in writing performance assessment, but concludes that no matter how detailed and thorough it is, there will always be discrepancies between raters in their decisions.

Continuous monitoring is another issue in rating exercise, as the effect of training can fade with time, so in order to maintain consistence in rating it is desirable to insert regular training sessions in the rating process especially in large-scale assessment procedures. Congdon and McQueen (2000) in examining rater stability found differences in rater severity from day to day and between the beginning and the end of the rating exercise. They suggest that monitoring of raters is needed with continuous retraining and checking for accuracy in rating (2000, pp. 175-176). However, gaining experience in rating can result in differences in rater severity and leniency. Shaw (2002a) examined the effects of training and standardisation iterations within a group of trained examiners and found that raters' assessing performance on the fifth iteration was the most lenient. He attributes this change to raters' lessening of confidence in consistency of rating decisions.

As training raters seems to be both time-consuming and costly, a solution can be to utilise modern technology providing online training with an appropriate computer programme. Computerised online training programme has several advantages compared to live training sessions, such as raters can work at their own pace, they can work whenever convenient and can work alone (Elder, Barkhuizen, Knoch, & Randow, 2007, p. 50). A study designed to examine whether rater agreement and consistency improve following training on-line has found that raters' online training is more effective than live training. Although results do not show considerable differences among raters, the overall impact on rating is beneficial and it is worth elaborating the training programme and implementing it in written performance assessment (Elder et al., 2007).

The above discussion shows that most authors agree on the importance of training, but there are other factors that should be considered which can result in inconsistency. Rater training is not the ultimate solution to explain the differences between raters, no matter how detailed and careful rater training is; there can still be some disagreement. "While rater training is essential for

creating the conditions for an orderly measurement process based on ratings by making raters more self-consistent, there is a limit to how far this process can be successful, or whether the elimination of differences is desirable!" (McNamara, 1996, p. 127). Raters' discussions during the standardisation meeting can reveal more about the way they make their rating decisions. They attend to different features of scripts and sometimes deviate from rating criteria. That is why it is important to control the discussion in rater training sessions and attempt to reach consensus by "positive sharing" (Hamp-Lyons, 2007). It follows that further research is needed into the way raters go through the rating process and arrive at a score. The score awarded as the result of assessment should reflect the candidate's performance. It is the product of several factors of the rating processes, such as "the prompt, the raters, and the scoring criteria and procedures" (Alderson & Banerjee, 2002, p. 98). The present study attempts to shed light on the rating processes of raters with similar characteristics and training to reveal how raters assess written performance and to find out more about the issues raised in connection with rater variables.

4.7 Conclusion

The chapter attempted to provide a deeper insight into the nature of rating process in written performance assessment and intended to highlight the issues of rating written performance that is dealt with in this study. First, the most commonly used rating instruments were presented and a comparison was made between holistic and analytic scales. There is no scale that is more appropriate than another, it is always the testing context that determines which scale to use in scoring procedures. Raters play a distinguished role in the rating process, as the score awarded is their decision that they make on the basis of the script and the rating scale. Raters' interpretation of scripts and rating criteria show great diversity (Wolfe, 1997), so the way they perceive the elements of the rating process influences the judgements they make. In addition, rater characteristics, such as their background knowledge, or their "expertise" in rating is also decisive. In a study to improve rater training, Hamp-Lyons recorded standardisation sessions the analysis of which showed that raters develop individual approach to rating and verbalise their judgements diversely (Hamp-Lyons, 2007, p. 4).

Rating processes can be viewed from different aspects, they can be perceived as problem-solving activities, or as a set of different strategies. The review of the five most recent frameworks shows the complexity of the processes raters go through when rating written performance. Rater agreement can be looked at from different aspects and some kind of compensation for inconsistency is needed. Training raters is necessary to reach a common ground

in rating, however, it is not the ultimate solution for maintaining reliability of scoring. Raters should be monitored continuously regardless of experience they have in rating. Research up to date concludes that further investigation into rating processes is justified to reveal more about the features of the processes raters go though when evaluating written performance. The present study attempts to contribute to this research area and examine rater behaviour in Hungarian EFL context.

Chapter 5

Verbal Protocol Analysis as a Research Method in Written Performance Assessment

Introduction

Before presenting the study and identifying the research questions, I introduce think-aloud protocol as a research method. First, I present the main approaches to research in L2 acquisition. Then, categorisation and nature of verbal reports follows. Finally, I discuss the use of think-aloud protocols in written performance assessment research together with the stages of implementing the procedure.

The most appropriate technique for following mental processes is borrowed from psychology and is called introspection, which means the observation of mental processes via analysing verbalised thoughts. Contrary to concerns that some researchers share in connection with verbal data as a valid base for empirical research, Ericsson and Simon (1993, p. 9) argue that observation of cognitive behaviour results in as valid data as observation of any other type of behaviour. However, verbal protocol analysis methodology has several features that need careful consideration to ensure validity and reliability.

The application of verbal protocol analysis in language testing research has become widespread recently. Areas of research cover assessment of the four language skills: reading, writing, listening and speaking, and translation (Elekes, 2000; Leow & Morgan-Short, 2004; Nikolov, 2006). Research can focus on test takers examining their cognition and reveal how they approach the test and what strategies they use (Cohen, 1984).

Written performance assessment research focuses on several aspects: it can be geared towards the composing processes or can deal with skills needed for written production, and thinking processes involved or features assessors focus on. In written performance assessment raters' role is seen as one of the three sources of measurement error, the other two sources are issues related to reliability and construct validity (Milanovic et al., 1996, p. 93). As discussed in Chapters Three and Four, the focus has shifted to the way raters make their decisions due to the growing interest in performance assessment in which interplay between several factors including the raters' role is decisive (Cumming et al., 2002; DeRemer, 1998; Lumley, 2000; Milanovic et al., 1996;

Wolfe, 1997). These are only some of the areas that written performance assessment research deals with. In order to be able to answer any questions triggered by different aspects of written performance assessment research has to be designed carefully. Research in the present study focuses on raters' decision making and aims at following their thinking. This chapter presents those issues related to verbal protocol analysis as research methodology that are relevant to the present research.

On the one hand, the choice of verbal protocol analysis as research methodology in this study is justified by its frequent use in similar studies. On the other hand, as verbal protocol analysis has not been employed widely in Hungarian L2 research, this study attempts to raise researchers' interest in verbal protocol analysis. Elekes (2000) conducted a study into readers' thinking and claims that verbal report as a method is a highly effective way of gaining information on participants' cognition. Nikolov (2006) followed Hungarian students' strategy use during taking reading and writing tests. This chapter provides an insight into verbal protocol analysis drawing attention to the most relevant issues for the present study.

5.1 Research in L2 Acquisition Studies

Research in L2 acquisition, as in any field of science involves finding answers to questions or verifying hypotheses that emerge while studying it. Methods used in research should be falsifiable and replicable as far as the issue of truth is concerned. If a statement resulting from any research cannot be falsified, the research cannot be scientific. Similarly, research should be replicable, which means that if it cannot be repeated, it is not scientific (Gass & MacKey, 2000, pp. 3-4).

Two broad types of research are distinguished: quantitative and qualitative one. Quantitative research is based on numerical analysis of data to verify a hypothesis, whereas qualitative research is interpretive, it is not easy to express and analyse data numerically (Mackey & Gass, 2005, p. 2). Thus, research data can be quantitative and qualitative, and are collected either by experimenting or observation, and analysed employing statistical or interpretive means. Data are collected in quantitative research with means of controlled measurement, while in qualitative using some kind of controlled observation.

Data called "hard" and are collected in case of quantitative type of research and they serve as a basis for generalisations. "Soft" data are collected using qualitative research, which are not easily generalisable, they rather show tendencies and processes (Mackey & Gass, 2005, pp. 2-5). This distinction is not as straightforward as it seems and as Mackey and Gass emphasise, the different research types can be placed on a continuum with several possible

variations (2005, p. 2). L2 acquisition research is characterised by using both, in many cases within one study to triangulate the findings.

Second language acquisition research is multidimensional in character, research methods depend on features of inquiry and research is designed according to research questions. It follows that there is no uniform model for data collection procedures, and data collection should result in analysable data (Mackey & Gass, 2005, pp. 43-46). Research usually requires data from several sources, which are analysed and conclusions are drawn based on both of them. Triangulation also involves different data collection procedures, but the aim of triangulation is to provide evidence for the relevance of findings and results.

Quantitative research data can be collected without intervening in the process, participants act as they would in normal operational circumstances. This suggests that data obtained this way mirror the actual phenomenon and analysing them the researcher can answer the research questions or support the hypothesis.

However, especially in the field of social science to which applied linguistics belongs, research questions cannot be always answered by mere statistical analyses of quantitative data, as interest is not only in statistical qualities of data, but in the way these data are generated. Qualitative data collection procedures include introspective measures with which it is possible to get insight into what goes on in participants' minds when performing certain tasks (Gass & Mackey, 2007).

5.2 Introspective Reports as Data Collection

Introspection as a method for gaining insight into thinking processes of participants originates from psychology and is based on the assumption that mental processes can be observed similarly to external processes. Participants in research using the method are asked to verbalise their thoughts during task completion or problem-solving. Introspective reports are of several types and differ according to the time frame, form, task type and support provided by the researcher (Mackey & Gass, 2005, p. 77). Popularity of introspective methods in L2 research is due to possibilities of getting an insight into participants' language formulation and into the thinking processes they go through (Gass & Mackey, 2000, p. xi). Traditionally verbal protocol data have been considered as "soft", as data do not come from direct observation. The advance of technology makes it possible to record cognitive data more precisely thus making them "hard" (Ericsson & Simon, 1993, pp. 3-5).

Verbal protocol analysis is often used in L2 research. Although the method is closely related to the methodology of introspection and the

terms introspection and verbal reports are sometimes used interchangeably, there is a fundamental difference between the two. Verbal protocol analysis differs from the way introspection is applied in psychology, as it requires thought verbalisation, whereas introspection builds on inferences made by individuals. In verbal protocol analysis inferences on thinking are provided by the researcher after analysing data. Verbal protocols constitute data for analysis and the method is considered qualitative in nature. However, there are several possibilities for quantifying and analysing data from quantitative aspects. It follows that verbal protocol analysis can be applied as one of the methods in L2 research or as research in its own right (Green, 1998, pp. 1-4).

The theoretical relevance of cognitive science in think-aloud protocol analysis, especially in the field of assessment is apparent, as it involves decision-making processes which are judgements. Although research reported by Greatorex and Suto and discussed in Chapter Four (Greatorex & Suto 2005; 2006) is based on GCSE marking processes, most findings can be related to written language performance assessment as well. Using think-aloud protocol analysis they investigated raters' decision making processes from a cognitive perspective and identified five cognitive strategies that raters employ in marking GCSE papers. The five strategies can be either of System 1 or System 2 type. System 1 type strategies are those thought processes which take less time and are associative. Matching, scanning and no response strategies belong to this group as they comprise simple pattern recognition to identify whether the answer is acceptable or not. Evaluating and scrutinising strategies belong to System 2 type thought processes. Evaluation strategy involves judgement about the candidate's answer, the examiner relies on knowledge and information s/he has and decides whether further strategy is needed. Scrutinising involves an examiner's reconstructing a candidate's intention in case the answer is not usual or if it does not confine the scoring rubric (Greatorex & Suto, 2006, pp. 7-13). It follows that interpretation of raters' thought processes in decision-making is highly relevant from cognitive aspect as well.

5.3 Categorisation of Verbal Protocols

Verbal reports are considered as a special form of introspection and can be categorised from different aspects. Cohen (1994a, p. 679) makes a distinction between three main categories: self-report, which is general information about something; self-observation, which is a report to explain what someone is doing or has done; and self-revelation, which constitutes a report on thinking and is also called think-aloud. During self-observation participants provide explanations of their actions, thus such reports reflect their reasoning; while

in self-revelation participants verbalise their thoughts directly. The main value of the method, according to him is its immediate feature, i.e. verbalisation happens during or very close to the activity observed.

As there are different types of verbal protocols, the choice of the most suitable data collection method is the researcher's responsibility and it depends on the research question. For research in language testing Green (1998, pp. 4-7) proposes a taxonomy similar to Cohen's. She provides an explanation for the elements of the three dimensions of the categorisation.

First, the form of report has to be determined, which can either be talk-aloud or think-aloud. Although the difference between the two is very subtle, they are not identical. Talk-aloud covers eliciting what is already in verbalised form in mind, whereas in think-aloud information transformation is also involved, as it attempts to elicit non-verbal information.

There is a temporal dimension to refer to the time of report, to identify whether it appears simultaneously with the elicitation procedure or after completing the task. In this respect both talk aloud and think-aloud techniques can be concurrent or retrospective. Conrurrent technique involves verbalising thoughts during task accomplishment, which can be demanding as the participant has to concentrate on the task and on thought verbalisation at the same time. Retrospective data are collected after task completion and they should be dealt with great care as with time information can be distorted and altered. That is why retrospection is recommended in cases where concurrent reporting is not possible, such as during oral interview tasks.

Procedural variations refer to the type and amount of prompting or mediation provided during task accomplishment. The researcher can mediate in each type of verbalisation depending on the research type, but it is important to consider that thought processes can be influenced by researcher's intervention (Green, 1998, pp. 9-11). This categorisation is similar to the one Gass and Mackey propose who have an additional category, the form of the report, which means that the report can appear in an oral or written form, or in both forms (Gass & Mackey, 2000, pp. 13-15).

5.4 Concurrent Think-Aloud Protocols in Written Performance Assessment

Within research into assessment of writing skills, there are two main focuses: research into the processes of composing processes and research into the processes of assessment of written texts. Research into assessment of written performance attempts to shed light on the issues of construct validity, scoring procedures involving measurement instrument and rater behaviour, and

considerable effort has been devoted to compiling a model of the rating process (Cumming et al., 2002, pp. 67-68). Cumming et al. present three consecutive studies for gradually building a framework for written performance assessment empirically by following raters' decision making processes to reveal what they attend to when rating. Each of the three studies is built on concurrent think-aloud protocol analysis to supplement other data in following raters' decision-making processes (Cumming et al., 2002).

As categorisations above show, several types of verbal protocols are used depending on the type of task which is the basis of observation and research questions. One of them, think-aloud protocol analysis has been employed to investigate performance rating procedure from several aspects. Performance rating can be perceived as a problem-solving activity, which thus can be looked at by analysing raters' thinking. The analysis of raters' thinking or decision-making focuses on several aspects of the rating processes: the role of the assessment criteria and the way raters interpret them; raters' characteristics including their background and expertise; and the effect of rater training on the rating process. In addition, the rating procedure has been examined, especially the way raters arrive at a score and strategies they use during rating (Lumley, 2000, pp. 24-39).

Written performance assessment involves problem solving activities (DeRemer, 1998; Mackey & Gass, 2005, pp. 79-85), so concurrent think-aloud protocols seem to be relevant to make inferences on raters' decision making processes and this way enhance validity and reliability of rating. DeRemer argues that written performance assessment has the features of ill-structured problems as it is more complex, less definite, and there is no single "good" answer to the problem (DeRemer, 2002, pp. 13-14). It follows that examining raters' cognitive processes is justified, as raters' decision making is not a simple comparison, but a complex problem solving process which is different from rater to rater. Findings of think-aloud protocol analysis of the study revealed that raters approach the rating task from two aspects: they focus on the text or on the scoring criteria. DeRemer remarks that both text-based and rubric-based approach to the rating task have their consequences on validity of scoring as the meaning of the awarded score is not necessarily the same.

Two types of validity and reliability issues are related to verbal protocols: one comprises validity and reliability of the data collection technique, and the other concerns validity and reliability of coded data. Both of them can be sources of measurement error and they contribute to the success of research. Validity of the technique refers to the extent to which the verbalised information corresponds to what is actually thought by the participants without distortion. The issue of reactivity, which is "the act of thinking aloud potentially triggering changes in learners' cognitive processes while performing the task" (Leow & Morgan-Short, 2004, p. 35) has generated some investigations. Effects of thinking aloud during reading and controlled writing have been compared to accomplishing similar

tasks without thinking aloud. Leow and Morgan-Short concluded that thinking aloud does not have distorting effect on either text comprehension or controlled writing. These findings suggest that readers can provide informative data on their cognitive processes, which justifies the present research into raters' thinking processes while rating written performance.

In order to maintain validity of concurrent verbal protocols it is important to provide participants with appropriate instructions for carrying out the task and make clear that they are not expected to explain what they think or do. The validity of the technique can also be affected by researcher intervention, the less the intervention is, the less likely the distortion of data. Reliability of the technique depends on the validity of the coding scheme. Individuals produce different protocols as they think differently. Therefore, in order to be able to describe certain behaviour reliably, protocols should be coded accurately to capture the cognitive processes.

Apart from the validity and reliability of the data collection technique, the validity and reliability of the coded data have to be considered. The validity of the coded data as mentioned above is related to the accuracy of capturing thinking processes. The assumption is that if two coders were asked to come up with a coding scheme for the same protocols, their coding categories would be similar. Maintaining reliability of coded data involves either employing a second coder or coding the data twice. These concerns of validity and reliability are inevitable in any language testing research using any type of verbal protocol analysis (Green, 1998, pp. 10-14).

5.5 Verbal Protocol Analysis Procedure

Application of any research methodology, including verbal protocol analysis, needs careful consideration as far as its relevance to the research is concerned. As mentioned earlier, verbal protocol as data collection measure can be employed on its own in language research, but most often this method is combined with other techniques to complement research. Regardless whether it is used alone or in combination with other data collection methods, there are certain steps that should be followed in research design.

There are three main stages of verbal protocol analysis: data preparation and collection, coding scheme development and data analysis. Data preparation starts with task identification, "task in this context refers to an activity that may be carried out by an individual, perhaps a test taker or an examiner" (Green, 1998, p. 35). It means that the most appropriate procedures have to be chosen for eliciting as much and as accurate data as possible. Then, task analysis follows which is defining in advance what participants are likely to do and how they are going to behave. It also helps to set up initial

categories for the coding scheme. Next, the procedure is chosen on the basis of whether data are collected using think- or talk-aloud procedure, whether it is concurrent or retrospective, and whether intervention is needed and if yes, how much. Participant selection involves choosing from a sample those who are representative of the population in question. Then, they need training in order to maintain consistency in task completion. Verbal data can be collected using different methods, such as note-taking, tape or video recording. The most reliable means are those which allow for collecting all data for analysis. Verbal data can be supplemented with additional information from different sources to produce a wider perspective for research (Green, 1998).

The first step in processing the collected data and compiling a coding scheme is transcribing the protocols. There are several possibilities; the researcher or professional transcribers can perform transcription. The transcripts are then appropriately segmented and finally coded. The coded protocols are ready for analysis after reliability check, which is mentioned above, and can involve either a second coder or coding the protocols twice (Green, 1998, pp. 14-21). The steps that are most relevant to the present study are discussed in more detail in the next sections.

5.5.1 Data Preparation and Collection Procedures

Data preparation procedure starts with task identification and task analysis. The researcher decides what steps are involved in accomplishing the task, in case of rater decision-making process research the main strategies and steps raters are likely to follow can be established a priori and these serve as a basis for compiling the coding scheme. This type of research is called data driven as not all categories for analysis can be established a priori; the complexity of the task requires modification or amendment of the procedure. An illustration of such a data driven study is presented by Milanovic, et al. (1996) in which a tentative model for the marking process is set up initially as a result of deciding in advance how markers are likely to proceed in rating (Green, 1998, pp. 35-40). It shows that thorough task analysis reveals the processes participants go through during task accomplishment, which helps the researcher to identify categories for the coding scheme.

The following phase in data preparation is choosing the appropriate data collection procedure. In order to get as informative data as possible, the most recommended method is concurrent protocol, which means that participants verbalise their thoughts as soon as they occur in mind and this way little distortion is possible. Concurrent data collection seems to be highly relevant in written performance research, especially in examining raters' marking processes, as the procedure does not substantially interfere with rating (Green, 1998, p. 40).

5.5.2 Verbal Report Procedure Preparation

When the task is identified and the data collection procedure selected, the procedure for task accomplishment is designed and prepared. It means that considering the protocol type, instructions for participants are needed in which they are informed about the task, the equipment for recording. It is also useful if participants have the opportunity to try the procedure out, especially if they have never had such an experience before. Preparation is inevitable for ensuring validity of the task as any misinterpretation results in inappropriate data which should be discarded later. Preparation consists of several steps: instructions, practice, while and post sessions (Green, 1998).

A verbal report is not a social act, participants have to be informed properly what is expected from them and they have to focus on task and not on thinking aloud (Ericsson & Simon, 1993, xiv). Instructions for them include all necessary information needed for verbal report production. The researcher should explain briefly the aim of the study without outgiving too much about the research questions which can bias participants. It is also important to make clear what is required from participants, for example, in the case of think-aloud concurrent protocol they are expected to verbalise each thought that comes to their mind while carrying out the task. It is inevitable to make evident that they are not expected to provide explanation or make inferences about their thoughts. That is why practice is needed during which a similar task is assigned to check that the participants can accomplish the task as required. In addition, the equipment for recording should be tried out to see that participants are not influenced in any way by being recorded. If the researcher is present during the task, decision should be made on the amount and type of prompting if necessary. After data collection the researcher can conduct a debriefing session (Green, 1998, pp. 41-50).

5.5.3 Data Transcription

Data collected are presented in form of notes, tape or video recordings and first they need to be converted into written form. Then, considering the research questions they are segmented and coded. So data collected have to be transcribed to make it suitable for further processing. This phase is the most time consuming in any research and usually professional audio typists are involved. However, it can be useful for the researcher to transcribe at least a portion of the protocol or listen to some of the recordings. No matter how precise the transcription is, it cannot duplicate all features of audio or video recording. It follows that transcripts should contain recorded information accurately without any alteration, even if there are incomplete sentences, false starts, or grammar errors in the text (Green, 1998, pp. 50-66).

5.5.4 Segmenting the Protocols for Coding

When the transcription of the collected data is completed, the following phase is segmenting, that is making them ready for coding and later for analysis. Data can be coded depending on the measurement instrument, which can be a nominal, an ordinal or an interval scale, each of which lends itself to examining data from different aspects. A nominal scale provides information on the type of data, for example if the participants are male or female, or if the think-aloud protocols are in the target, native or mixed language. An ordinal scale shows ranking of the data, this way we can distinguish between low, intermediate or high achievers in a test. In addition to rank order, an interval scale gives information on the distance between variables as well (Nahalka, 1996, pp. 350-352).

Verbal data, especially oral ones, need to be prepared for analysis; they cannot be directly assigned to any coding scheme. In many cases L2 research data are gathered from oral sources, which need some kind of transformation before analysis. Data transformation, as mentioned before, means transcribing the data first, which is followed by segmenting the protocols before coding. Segmentation means dividing the written form of verbal protocols into manageable units and preparing them for coding. There is no general rule as far as the length of each segment is concerned. A segment can be a single word, a phrase or a clause, in some cases even longer bits of texts can be identified as text units of a protocol. The most important criterion for segmentation is that each text unit should represent an activity even if it is an incomplete utterance (Green, 1998, pp. 73-78).

However, segmentation is not without dilemmas and decision-making, Lumley (2000, p. 125; 2002, p. 253) calls segmentation an arbitrary process. The principles of protocol segmentation in his study are based on the research questions. It means that text units are identified according to the research questions and not text boundaries. The aim of his study is to examine raters' rating sequences, their criteria in making decisions, problems in rating and raters' comments.

5.5.5 Developing a Coding Scheme

After the segmentation of verbal protocols is accomplished, the segments have to be made identifiable for analysis, that is, they have to be coded. There is no uniform method or pattern for assigning codes to segments; researchers have to develop their coding schemes for their study: categories should fit the data and the research questions. However, there are certain guiding principles which need consideration in order to maintain the validity and reliability of coding. Employing broad categories end in high reliability but less informative data, whereas too many details have negative effect on reliability. If during

coding or analysis it turns out that categories are too broad or difficult to handle, they can be aggregated into less and more manageable categories. The researcher has to aim at a balance in deciding on codes, if there are more details than needed, coding categories can be combined during data analysis, but it cannot be done the other way round. Most often coding scheme development is preceded by data collection, in such a data driven study the general categories are refined.

Coding schemes can be developed on the basis of a theoretical framework in advance, but they also might need alterations (Green, 1998, pp. 68-71). As Green puts it, "the main purpose of a coding scheme then is to capture as much information within the set of protocols as possible. Coding categories should be quite specific. At the same time, though, coding categories must be generalisable across a number of protocols..." (1998, p. 70). It follows that coding scheme development involves dilemmas that need careful consideration. As there is no unified method for coding verbal protocols, first a tentative coding scheme can be compiled based on a theoretical framework, and then, bearing in mind the research questions further categories may be added.

5.5.6 Analysis of the Coded Data

Verbal protocol analysis involves several different procedures, first of all reliability of coding is calculated. Reliability of coding can be checked by either employing two coders or coding a portion or all protocols twice, a combination of the two is also possible. Inter-coder reliability shows agreement between coders; intra-coder reliability estimates show consistency of a coder: both indicate the credibility of a study. Further data analysis techniques depend on the research questions: there are many ways to analyse data, either by hand or using software packages (Green, 1998, pp. 12-13). In order to maintain data reliability and validity, before the operational stage a pilot study is usually conducted in order to find out whether materials and methods intended for research are appropriate for the given research. A pilot study has a narrower focus and involves fewer participants (Mackey & Gass, 2005, pp. 43-44).

The pilot study for the present research, as presented in Chapter Six, attempted to produce a coding scheme for following five raters' thinking during rating written performance (Bukta, 2007). Findings of the pilot study served as a starting point for developing the coding scheme for the present study. In addition, several dilemmas have been generated regarding interpretation and analysis of verbal data. For example, the question of language of the protocols has caused some problems, as the raters are non-native speakers of English and they rated English scripts using a rating scale in Hungarian, but they were not required to use English exclusively in their think-aloud task, which resulted in three types of protocols: English, Hungarian and mixed language.

5.6 Advantages and Limitations of Verbal Protocol Analysis as Research Methodology

Verbal protocol analysis as a research methodology in L2 research has gained significant interest recently. However, as any methodology it has its advantages and disadvantages which have to be taken into consideration when designing research.

The first question is whether the protocols reflect thinking processes without being influenced by the fact that they are recorded. Participants can report their thoughts inaccurately due to unconsciousness of their thinking or lack of appropriate verbal skills. At the same time, verbal reports reflect thinking closely, so it is possible to follow thinking (Gass & Mackey, 2000, pp. 105-107). The issue of reactivity or the influence caused by thinking aloud on data has been investigated in a reading comprehension research context. Leow and Morgan-Short (2004) provide empirical evidence as a result of comparison of two groups accomplishing a reading comprehension task, one with think-aloud, the other with non think-aloud procedure. They conclude that thinking aloud does not interfere with cognition and it does not influence participants' ability in task performance (Leow & Morgan-Short, 2004, p. 48).

Ericsson and Simon deal with the concern raised by behaviourists who say that cognition is different if thoughts are verbalised and thus verbalised and not verbalised thinking are two independent processes. They argue that verbalisation is a true reflection of what goes on in one's mind thus inferences can be made about cognition (Ericsson & Simon, 1993, pp. 2-8). There are two other characteristics of verbal protocol analysis which need careful consideration when the method is employed. One is the idiosyncratic feature of verbal reports: each individual is different in cognition and they can give an account of their thinking in several ways. The other characteristic feature is related to the previous: if verbalisations are diverse, their encoding is not straightforward which can lead to difficulties when making generalisations. These issues can be solved by accurate and careful data preparation (Ericsson & Simon, 1993, pp. 169-170).

The influence of methodology on research is discussed by Lumley when he lists his observations on rater behaviour. His study aims at examining rater behaviour when marking written performance. They are experienced in rating, but not in providing concurrent think-aloud protocol during the rating task. Some of them reported differences in the way they accomplished the rating task compared to operational rating: one appeared to be more organised, others less. There was a rater who felt less comfortable than during an operational rating. He concludes that raters' cognition is much more complex than what they can verbalise. They cannot give an account of each thought that goes through

their mind. However, verbal protocol analysis contributes to research in rater behaviour, especially in finding out more about strategies raters employ (2000, pp. 282-286). He says that the think-aloud "requirement appears to push this basic simplifying process still further, but the reliability of the scores obtained suggests it does not completely derail it" (Lumley, 2000, p. 285).

Think-aloud protocol analysis takes considerably longer than any other data collection methodology, as data processing including transcription, segmentation and coding the protocols before analysis is a time-consuming activity. In order to maintain credibility of such study, data processing should be accomplished in a rigorous manner. Thus, smaller sample is considered, which has an effect on statistical analysis, statistical differences are not as apparent as in large-scale studies. Wolfe suggests either using a larger sample of participants or simplifying data processing methodology (Wolfe, 1997, p. 103).

To cater for some of the inadequacies of verbal reports mentioned above, researchers employ more than one methodology for data collection. Milanovic, et al. designed the first phase of their research for following holistic markers' decision making using two types of verbal reports and a group interview. Participants were asked to provide retrospective report after each composition they rated, but only part of the compositions were rated using concurrent verbal report. They say that a single verbal report may not provide accurate data of the cognitive processes, so they find it justifiable to use a combination of retrospective written report, introspective verbal report and group interview for collecting valid and reliable data, as data collected with different methodologies can contribute to the credibility of the research (Milanovic et al., 1996).

5.7 Conclusion

The chapter attempted to locate verbal protocol analysis as a research methodology within the field of research in L2 acquisition. First, I explained the notion of verbal protocol and its relationship with psychology. Then, I presented the need and ways of establishing and maintaining validity and reliability. The main issues in verbal protocol analysis were discussed with special attention to its application in written performance assessment research. Written performance assessment is characterised with rater, rating process, rating scale and candidate variables; thus, rater behaviour is in the centre of attention of researchers. Tracing the way raters arrive at a score is an intriguing issue. As Lumley puts it, "it is clear that we only get fragments of what raters think, but there is still no reason to suppose that they did not think what they describe" (Lumley, 2000, p. 66). We can get valuable data from raters' verbalisations. Each research employing think-aloud protocols is different, but data collection methodology and ways of data processing are characterised by distinctive principles based

on firm theoretical considerations. Data analysis can only be conducted credibly if data are prepared following careful consideration of research aims. Central to verbal protocol analysis is segmentation and coding the protocols to arrive at a statistically analysable dataset. Finally, I listed some limitations together with the merits of the research methodology.

PART II

Investigating Raters' Decision-Making Processes and Awarded Scores in Rating Hungarian Efl Learners' Compositions

In the second part of the book, I present the pilot study (Chapter Six) which aimed at tracing raters' decision-making processes. The following chapters, from Chapter Seven to Chapter Ten, include the research conducted into raters' decision-making processes.

Chapter 6

A Pilot Study: Tracing Raters' Decision-Making Processes

Introduction

In Chapter Six I would like to present a pilot study that aimed at investigating raters' decision-making processes. Raters' role in written performance assessment is decisive since the raters are one of the sources of measurement error. This is why it is important to have an insight into their thinking during rating. The aim of the pilot study is to trace Hungarian raters' decision-making process and find out about the relationship between the performance, the scale, and the rater (Bukta, 2007). First, I will explain the background of the study. Then, I will introduce the research questions and the research design. I collected data on raters' thinking processes in order to get an insight into their thinking processes. The data on raters' decision-making were categorised with a coding scheme whose development is detailed. Finally, I will present the analysis of the data and the conclusion I drew. The intention is to understand better what features of written performance raters attend to when marking compositions and, most importantly, how these rating processes can be observed.

6.1 Background to the Pilot Study on Assessment of Written Performance

The study attempts to shed light on the role raters play in written performance assessment. The study was carried out using the data collected in spring 2003, during the national survey of learners' language abilities conducted by National Public Education Testing Centre in Hungary (Országos Közokotatási és Értékelési Vizsgaközpont). The writing task and the analytic scale are part of the training material and are not produced by the author of the present study (Nikolov & Józsa, 2003). In Hungary student performance assessment in different school subjects has been carried out for years and the findings are valuable sources for policy-makers, teachers and researchers alike (Csapó, 2002). The survey in spring 2003 intended to assess language learners' performance in the two most popular foreign languages taught in

Hungarian schools, German and English. Two age groups were considered: a representative sample of the student population in the 6th and 10th years. The survey was carried out in three language skills: reading and listening comprehension, and writing. Testing the skill of speaking was excluded, as it would have been costly to organise for such a large population. The only productive skill evaluated was writing. The pilot study focuses on the raters' assessment processes of students' written performance in the 10th year using analytic scales.

Marking written performance using rating scales involves raters' decision-making, which is influenced by their "assumptions, expectations, preferred rhetorical models, world knowledge, biases, and notions of correctness" (Cohen, 1994, p. 308). It follows that the raters' thinking process plays a significant role and should be considered carefully, as the score they arrive at is the result of their understanding of the performance, the assessment criteria, and the task (Hamp-Lyons, 1990). Furthermore, raters can be influenced by their expectations in connection with the task, the candidate and they can attend to surface characteristics of the compositions as well (Weigle, 2002). It is essential to make sure that raters can apply the criteria defined in the scale consistently and make their decisions excluding any subjective judgement.

That is why it is inevitable to standardise the rating process and make sure that raters can focus on the performance only and that other features of the scripts do not influence them. To ensure consistency in judgement, rater training is essential prior to the assessment procedure, during which raters should be familiarized with the test construct, the task, the scale, and the way they can arrive at a decision (Alderson, Clapham, & Wall, 1995). Thus, the assessment of written performance is affected by several factors among which the raters' decisions play a significant role. No matter how detailed and substantial the task design, the scale and the training are, there can still be some aspects of the rating procedure that need attention. As raters have different experience and background they may vary in decisions, they may focus on one criterion more than the other (Cumming, Kantor, & Powers, 2002).

The nature of the evaluation process does not allow direct observation, as thinking is involved, so it is difficult to design a research instrument for the decision-making process without interfering with the assessment procedure itself. In order to collect data on raters' marking, they can be asked to think aloud during the assessment process and their utterances are audio recorded. Then, the recordings are transcribed for the purposes of a detailed analysis. This type of research allows tentative generalisations only, as there are certain factors that influence the results, such as the raters' ability to concentrate on verbalisation, the transcriber's ability to interpret each utterance properly, or the analysis of the data (Green, 1998).

6.2 Research Questions

In the pilot study, the focus is on tracing raters' thinking and an attempt is made to get an insight into their behaviour and investigate the way they arrive at their decisions. Thus, the following research questions were put forward:

- What are the features that raters attend to when evaluating compositions?
- How closely can raters' decision-making processes be followed?

6.3 Research Design

The study aims to focus on raters' decision-making process during rating and tries to explore what factors influence their decisions. In order to obtain data from raters, they had to think aloud and record their speech using a tape recorder. The collected data were transcribed and an analysis of verbal protocols was conducted to compile a scheme, which would allow for categorizing the data. Although significant research has been conducted earlier to trace how raters make decisions, it turned out to be difficult to allocate their utterances. Finally, the categorized data were analysed to find out how raters make their decisions.

6.3.1 Participants: the Raters

Originally, the rating procedure for the nationwide survey involved seven raters, three were reading writing tasks in German and four in English. One of the English raters was me, the researcher. The researcher conducted the training and did not take part in the think-aloud protocol exercise for the present study, so as not to influence the results with the knowledge of the research questions. One of the English raters' tape recording was so poor that it was impossible to transcribe, thus her data got lost.

Finally, five raters' think-aloud protocols could be analysed in the present study. They are in-service teachers of English and German, two of the German raters are English major graduates as well. None of them had received training for marking or had taken part in assessment in a nationwide survey earlier. Similarly, they had not taken part in a think-aloud protocol project either. Raters' identification numbers are as follows: first English rater (EngR1), second English rater (EngR2), first German rater (GerR3), second German rater (GerR4) and third German rater (GerR5). They are referred to in the rest of the paper with these identification numbers.

6.3.2 Procedures for Data Collection

Rater training for the nationwide survey was organised in two towns: Pécs and Szeged, Hungary according to a centrally accepted standard procedure and a training pack, which included sample scripts and the scores in the Pécs training. In Szeged and Pécs, raters of both languages were trained together with the intention of arriving at a consensus in the process of assessment and making the comparison between the two language performances possible (Nikolov, personal communication on 23 May, 2003). The procedure for training was elaborated and used by the English Examination Reform Project team for assessment of similar writing tasks (Alderson, Nagy, & Öveges, 2000). The scripts were collected centrally and then they were delivered to the local centres for assessment. Raters in Szeged took part in a rater training session conducted by the researcher of the present study, who had taken part in the assessment of similar surveys earlier (Csapó, 2002). The task and the analytic scale were part of the assessment procedure and not produced by the author of this study. In order not to interfere in the rating process of the national survey, the rater training in Szeged was supplemented by an element at the end, which aimed at familiarising the raters with the rationale of the research and the raters were prepared for the think-aloud procedure.

The rater training for the national survey consisted of two parts: after introduction, considerable practice followed aiming at standardisation. The procedures included a brief summary of the principles in testing L2 and the rationale of the survey. Then, each rater assessed the same script, the German raters a German one and the English raters an English script, respectively. Next, the raters justified their evaluation. They repeated the procedure three more times with new scripts. The trainer who compiled the training pack for rater training had chosen the scripts in advance. There were top and poor performances and the pack contained some scripts, which were problematic for some reason. The rating exercise ended with the summary of the principles that raters were supposed to follow during marking. The following part of the session focused on technicalities: how they should carry out the rating task and what help was available in case there are further problems.

The rater training in Szeged was supplemented with preparation for the present research. The last phase of the training directly related to the research and raters practised how to produce verbal protocols. First, the raters were familiarised with the principles of the research and the research questions. It was emphasised that the main goal was to get as much information as possible on the decision-making process during rating. Then, the rationale of verbal protocols was introduced and raters tried the procedure out. Raters had an opportunity to try the think-aloud procedure one after the other and monitor each other while rating the samples. They were finally asked to produce

think-aloud protocols and record them on audiotape. As the national survey focused on performance assessment and there was no intention to interfere with the rating process itself, data collection was limited to producing think-aloud protocols lasting ten minutes at the beginning, in the middle, and at the end of the rating process.

The researcher transcribed and analysed the data; there were five protocols altogether: the protocols of three German and two English raters.

6.3.3 Test of Written Performance: the Task

There were 4,013 scripts written by 10th year students in English and German delivered to the Education Centre in Szeged, Hungary for assessment. They were written all over the country in different school-types: primary schools, secondary grammar and vocational schools included. Table 9.1 shows the distribution of the scripts that raters assessed in Szeged.

Table 6.1
Distribution of English and German Scripts

	German	English	Total
Number of scripts – 10th year	1,867	2,146	4,013

The figures show that the number of scripts in year 10 in English and German is similar; there were more in English, though; the difference is 279 scripts. The total number of scripts assessed by individual raters was between 550 and 600 in both languages.

The task for both languages was the same; it was a guided writing task to produce a letter for an Internet magazine about a "dream" holiday. The prompt comprised of six content points guiding the students. The word limit was in the instruction: learners had to produce a letter of about 150 words (see Appendix 6.1).

6.3.4 The Assessment Scale

The centrally devised rating scale had been piloted in earlier surveys, and it was the same for the English and German scripts and was written in Hungarian. The scale was an analytic one divided into four areas; the first criterion was achievement of the communicative goal, the second referred to the quality and

range of vocabulary, the third accuracy, and according to the fourth criterion, text organization was to be measured.

There were five bands in the scale, each of them contained a range of descriptors starting with 0, the only band, where one score, zero could be awarded; there were two scores allocated for the other bands, leaving the rater some scope for more detailed assessment. There were equally weighted scores for each criterion; the maximum score was 8 points, making up a total of 32 points. Each band was carefully worded and was a qualitative descriptor of the language area construct in question. However, the first aspect contained a quantitative descriptor, the number of content points covered, six altogether, which appeared very clearly in the rubrics of the task (see Appendix 6.2).

6.3.5 The Coding Scheme

In order to be able to follow the decisions made during the rating process I compiled a coding scheme (see Appendix 6.3). I transcribed the recordings focusing on the utterances only and ignored the time spent with assessment. Raters spoke mostly in Hungarian language and in the case of German scripts; they used Hungarian language in verbalising their thoughts. The utterances were segmented and calculated (see Appendix 6.4). There were some instances when German examples were cited, which the raters translated into Hungarian. In a few cases though, they read out examples in German, which were words whose meaning could be deduced. It did not hinder the analysis of the transcripts. To illustrate the transcripts a segment of one of the protocols is in Appendix 6.5.

6.4 Results and Discussion

The coding scheme developed gradually, the protocols were first segmented and then they were numbered and labelled. The utterances were in some cases complete sentences, some were incomplete ones, and there were mere one- or two-word remarks or just sounds indicating raters' approval and disapproval. The first draft of the coding scheme was based on the first protocol, which EngR1 produced. This resulted in preliminary categories, which were subsequently numbered, such as, "Identifies script", or "Rereads script", etc. Then, I put the transcript aside for a time and segmented the rest of the transcripts. Some time later, I reread the first transcript to check intra-rater reliability.

After that, the categories were grouped according to the topic of the utterance, thus there were eight criteria of assessment identified: scoring technicalities, reading the script, general comments, rater behaviour, communicative goal, richness of vocabulary, accuracy and spelling, and text organisation. Then, the utterances in the verbal protocols were categorised one by one, starting with

GerR3, then GerR4's transcript was labelled followed by EngR2's one, and last GerR5's protocol was considered. Some new categories were established which completed the coding scheme; at the end of the procedure, there was a total of 50 categories (see Appendix 6.4). That is why the codes consist of a letter referring to the category of a particular behaviour and a number, which identifies the behaviour within that aspect. Numbering was not subsequent, as some were assigned later when that category occurred while reading and labelling transcripts one by one.

When the coding scheme was completed, the analysis of the data followed. The five raters evaluated a different number of scripts, the German raters read significantly more, as there were several task sheets with no or very little language to assess. In addition, the English scripts were longer than the German ones.

Table 6.2
Length of the verbal protocols and the numbers of scripts evaluated with think-aloud procedure

Rater	Number of scripts evaluated with think-aloud procedure	Number of scripts with no or little language	Word number in all transcripts	Number of utterances in all transcripts	Mean word number in an utterance
EngR1	11	0	2,325	285	8
EngR2	10	0	2,530	148	17
GerR3	17	4	4,977	386	13
GerR4	36	16	4,760	445	11
GerR5	21	4	3,280	365	9

However, the analysis of the word number in the verbal protocols shows that German raters' protocols were longer than those of English raters, as is shown in Table 6.2; raters of German turned out to produce more language during the rating process.

The number of utterances is different as is the word number, however, the average word number shows that some raters used more language in an utterance. There are extremes within individual rater's transcripts, for example, EngR1 uses one word for finalising the score and uses 41 words to evaluate content points. On the other hand, the length of the statements depends on the assessment behaviour; there are one- or two-word-long technical remarks and personal reactions, the latter sometimes was just a sound.

6.4.1 Distribution of Comments During Rating

Raters' comments made it possible to trace what they attended to and how they arrived at a score. Raters practiced the rating procedure thoroughly during the rater training, and they tried to follow it closely, although in some instances, they deviated from the agreed pattern. In Table 9.3 there are the numbers of utterances and percentages for the eight categories that were identified during the rating process. The eight categories were further divided into subcategories and the number of comments raters pronounced in each subcategory is in Appendix 6.5

Table 6.3
Number of Utterances of Five Raters during the Rating Process

Category	Number of utterances	Percentage %
Rating technicalities	248	15
Reading the script	47	3
General comments	292	18
Rating comments	451	28
Communicative goal	213	13
Richness of vocabulary	102	6
Accuracy and spelling	87	5
Text organisation	189	12
Total	1,629	100

In the next sections, I discuss these categories one by one and illustrate each comment type with an example to investigate how raters arrived at a score. Raters' comments were translated into English and they appear in what follows.

6.4.2 Comments on Rating Technicalities

There were three types of comments in rating technicalities category and were labelled with letter "T" and a number. Raters were asked to identify each script by reading out the student's code on the task sheet (T13) and to nominate the scoring category on the rating scale clearly during the assessment (T21). There were instances, when they explained what exactly they were doing (T33), as in Excerpt 6.1.

Excerpt 6.1: An example of comment on rating technicalities

Rater ID	Rater talk	Code
EngR1	Script ID: 048015202	T13
EngR1	Now I am going to look at the number of content points	T21
GerR4	The scripts of the last round...	T33

Raters rarely indicated when they were reading the scripts first and were rereading them, and they did not read aloud the compositions. They sometimes said that they were reading through the text.

6.4.3 General Comments on the Scripts

First, after announcing the script number most raters assessed some surface features and made remarks on length, legibility and layout. There were altogether 292 general comments, which is 18% of all comments identified (see Table 6.3); 44 of them referred to layout and length. The eleven subcategories of general comments' codes contained a letter "G" and a number and behaviour types can be identified according to them in the examples. As the examples of rater language in Excerpt 6.1 show, raters talked about comprehension problems (G5), they expressed their initial impression either before the actual reading of the text or after the first reading (G10); handwriting played a role in rating (G11). Although not all raters verbalised when they were reading the scripts, it is clear from the protocols that most of them commented on length and layout (G15) before the first reading.

Excerpt 6.2: Examples of rater talk in the "General comments" category

Rater ID	Rater talk	Code
EngR1	It is very difficult to follow what is written	G5
GerR3	It is terrible, I would say	G10
EngR1	It is written in too small letters	G11
GerR3	Now, it is only three lines altogether, at best	G15
GerR4	Wow, s/he has misunderstood it	G17
EngR2	The composition is weaker	G18
GerR5	It doesn't turn out whether the letter is written to a stranger	G19
EngR2	It made me think that the candidate's language proficiency cannot be bad, expresses him/herself well, but the problem is as follows	G25
EngR2	Now the question is whether the rater should supplement the missing details according to her fantasy	G27
GerR4	Vocabulary can't be rich	G29
GeR5	I can imagine that they were sitting next to each other	G39

After reading the text for the first time, raters often expressed personal feelings (G17), mentioned script quality (G18) and remarked on relevance (G19). There were several comments made on the students' overall proficiency (G25), forecasting their language knowledge. Concluding remarks referred to rating (G27) and student's performance (G29). Some irrelevant notices on the circumstances, such as seating arrangement or possible cheating (G39) were observed.

6.4.4 The Way Raters Arrived at a Score

The comments uttered during the rating process mainly referred to the way raters arrived at a score, the total number of utterances in the "Rating comments" category was 451, which was 28% of all comments. They made their decisions based on careful consideration. In Excerpt 6.2 there are examples of rater talk during the decision making process. Codes of the ten subcategories contained letter "R" and a number, as in the examples in Excerpt 6.3. Most remarks related to the final score for each criterion of the rating scale (R9). The decision-making process was sometimes characterised by hesitations (R12); raters in 31 cases tried to find the most appropriate aspect on the scale to compare the script to (R14). Raters often summarised the rating process and explained the way they arrived at a score (R24) before announcing the final score (R26).

Excerpt 6.3: Examples of rater talk in the "Rating comments" category

Rater ID	Rater talk	Code
EngR1	So, I will give 6 points for this	R9
GerR5	So, how many points shall I give for this?	R12
EngR1	It is because there are no paragraphs, but there is logic in it	R14
GerR3	However, what he wrote is very little	R24
GerR4	This, this is worth 4 points	R26
EngR2	My other problem is that this envelope contains only very weak scripts	R28
GerR3	The previous one was similar and I gave a 1, so to be consistent, I am going to give a 1 for organisation. What did I give before? Yes, it was a similar performance. So, I gave a 1 there. No, to be consistent I am giving a 1 now.	R35
EngR2	And now I have to think hard and maybe, I will give a 2 instead of a 1	R36
GerR4	S/he understood the task properly	R37
GerR4	I think s/he wanted to say, to imply who s/he travelled with, so it is not there, but the point has been partly covered.	R38

There is an example of other influence (R28) when the rater talked about the script quality in one of the envelopes (scripts were collected in envelopes according to learning groups in schools). Raters sometimes compared the scripts to each other (R35), they reconsidered their decision (R36) and occasionally forecasted evaluation (R37). There were four instances when the reader completed the composition offering a solution to missing parts (R38).

6.4.5 Assessment of the Communicative Goal

Raters were expected to start evaluation with considering the communicative goal of the compositions. All of them spent more time with the assessment of the communicative goal than with the other three criteria, which shows in the number of utterances. The total number of utterances was 213, which is 13% of all contributions, out of which raters made 70 on the content. Excerpt 6.4 illustrates the way raters assessed communicative goal with an example of each rater behaviour type. The eight comment types in this category were coded with letters "CG" and a number. Apart from adding up the content points (CG2), raters cited from the scale (CG4), evaluated content points (CG6) and read out the text as an example (CG8).

Excerpt 6.4: Examples of rater talk in the "Communicative goal" category

Rater ID	Rater talk	Code
GerR3	Now, let's have a look at the number of content points so far ... it is five content points altogether	CG2
GerR3	3 or 4 points are covered appropriately	CG4
EngR1	S/he wrote what they did, wrote about the place, and wrote things that they did, but did not write how interesting it was for him/her. And, what s/he wanted to do next	CG6
GerR3	S/he says, "We are travelling to Hawaii, because the weather is nice. I am going with my friend".	CG8
EngR1	So, s/he wrote where, wrote about the place, about something that was interesting, that they had pizza; wrote about things they did, but did not write about the person s/he went with.	CG23
EngR2	It is striking that s/he writes about the given points comparatively well	CG31
GerR3	So, this "dream holiday", how can we understand that? Some students think that literally and write about a dream.	CG34
EngR1	No, if I have a look at the scale, there are six points covered.	CG41

They often summarised the content of the compositions to justify the score they awarded (CG23), evaluated communicative goal, making remarks on the quality of the content (CG31). They referred to the rubric (CG34) and compared the script to the scale (CG41).

6.4.6 Assessment of Vocabulary

Rating richness of vocabulary turned out to be straightforward, as the number of comments was significantly lower than for the other criteria of the analytic scale. The total number of comments for vocabulary assessment was 102, which is 6%. The four subcategories for vocabulary were identified with a letter "V" and a number. There are examples in Excerpt 6.5 for each behaviour type. When rating vocabulary, raters often cited from the scale (V4) and gave examples (V8). They also evaluated the aspect (V30) and compared the scripts to the scale (V41).

Excerpt 6.5: Examples of rater talk in the "Richness of vocabulary" category

Rater ID	Rater talk	Code
EngR1	This corresponds to the task; shows wide variety and selection, is appropriate	V4
GeR3	'Cat' is not absolutely relevant, but at least [s/he] can write it down correctly, 'Flug' for 'flying', s/he could write it as well, 'hotel room', s/he knows it also, 'strand', OK it is the same in Hungarian	V8
GeR4	Naturally, as the whole is very short, we cannot talk about rich vocabulary, but if I have a look at these four sentences, there are more, so there are more verbs used	V30
EngR1	Shows wide variety and selection	V41

6.4.7 Assessment of accuracy and spelling

Similarly to the rating criterion of vocabulary, raters' focus on accuracy was low: there were 87 remarks, 5% of all remarks, on accuracy. The identification of the four subcategories was "Gr" and a number. There were 65 comments out of 87 on grammar evaluation; raters did not provide examples, except in one case, and they did not often cite from the scale or compare it to the script. Excerpt 6.6 shows some examples of rater talk when evaluating accuracy.

Excerpt 6.6: Examples of rater talk in the "Accuracy and spelling" category

Rater ID	Rater talk	Code
GeR4	There are several basic mistakes, but the majority is comprehensible	Gr4
GerR3	The word 'train' is written correctly, but the verb 'travel' doesn't have the correct auxiliary	Gr8
EngR2	S/he is writing about favourite activities, and I think basic grammar is missing, there is no sentence without errors, the text because of basic structural errors is not comprehensible	Gr32
GerR4	I would put it into band 5-6	Gr41

In the first example the rater read out a scale descriptor (Gr4) and in the second there is the only example that a rater cited (Gr8). The third quotation from the protocols demonstrates how a rater evaluated the aspect (Gr32) and there is a comparison to the scale (Gr41).

6.4.8 Assessment of the Text Organisation

Finally, according to the rating scale, text organization was evaluated, the raters made 189 remarks (12% of the total). This aspect in the analytic scale contained several different text feature descriptors. As the results show, the raters mostly attended to text coherence (44 comments), to letter conventions (27 remarks), to paragraphing (37 remarks), and to sentence variety (48 remarks). Excerpt 6.7 shows illustrations of rater talk when evaluating text organisation.

Excerpt 6.7: Examples of rater talk in the "Text organization" category

Rater ID	Rater talk	Code
GerR4	Logical coherence is on the level of tenses, there are no jumps between present and future, or present and past, the sentences follow a chronological order, what happened to whom and when	O3
GerR5	So, it shows some letter characteristics.	O4
EngR1	There is no greeting and signature	O7
GerR3	"I am going to Hawaii as the weather is nice"	O8
GerR3	Paragraphing: there are no paragraphs at all	O16
EngR1	I'll have a look whether there are complex sentences, as I can see; there are no complex sentences here	O22
GerR3	It is something between 0 and 1	O40
EngR1	I am looking at the middle of the top band	O41

The first example is a rater's comment on coherence (O3); in the following another read a descriptor from the scale (O4); there is remark on letter conventions (O7); then, there is a sentence read from a script as an example (O8). There were also comments on paragraphing (O16) and on sentence variety (O22). The last two subcategories referred to evaluation of organisation (O40) and comparison of scripts to the scale (O41). These findings show that raters rather commented on the various text features than referred to the scale, they did not provide many examples or they did not often compare the scripts to the scale.

6.5 The Rating Process

Looking at each rater's decision-making process, there are some observable tendencies in the procedure they followed. EngR1 did not make any conclusion on students' overall proficiency, on rating, or on the performance in general. She also refrained from identifying other influence, comparing the scripts to each

other, or forecasting evaluation. It is also apparent that she rarely cited examples and followed the same rating process for all of the eleven scripts she evaluated. Excerpt 6.8 shows an extract from her rating process. First, she identified the script and explained what she was going to do (T13 and T33).

Excerpt 6.8: An example of EngR1's rating process

Rater talk	Code
Script number 12819113	T13
First, I am going to look at the letter. I'll check both sides of the paper to see how much s/he has written	T33
It is full.	G15
I think, s/he is going to write about everything.	G10
I am reading the letter and checking whether s/he has covered the content points.	Rd1
I can see that s/he wrote about where s/he had been, with who and how; s/he also said why.	CG6
Meanwhile I am checking accuracy.	T21
There is something interesting in it.	G19
S/he did not like the beach and did not want to come back.	CG23
The letter has an appropriate ending	O7
So, the first score is 8	R9
The letter is to a stranger and covers all content points	R14
Vocabulary is varied and more or less appropriate	V30

Then, she made comments on overall features (G15 and G10) and announced first reading (Rd1). Next, she evaluated the criteria in the rating scale: first content points (CG6), then, she announced another rating criterion (T21) and remarked on relevance (G19). She summarised text content (CG23) and commented on letter conventions before finalising a score (R9). After that she justified her judgement (R14) and moved on to evaluating vocabulary (V30).

The second rater, EngR2 evaluated ten scripts and most of them were problematic. She tried to find sufficient and appropriate language to evaluate, but sometimes it was not possible. Her protocol contains 18 remarks on relevance of the content to the task, she said, for example, "I think this information is absolutely irrelevant to the task". She often hesitated: "I don't know. I do not think it is acceptable".

One of the German raters, GerR3, had seventeen scripts to rate, four of which did not have sufficient language to evaluate. Her rating process was consistent, she explained thoroughly what she was doing and her protocol was the longest (see Table 6.2). First, she evaluated the layout, the length and comprehensibility and then she followed the rating scale starting with task achievement. She made

remarks on comprehension not only at the beginning. The rater sometimes referred to comprehension problems when evaluating the particular aspects, such as grammar: "The spelling is bad, but it is not impossible to make sense of it". The number of comments like this was 15 altogether and she expressed her personal reaction, for example she said, "That is very funny". GerR3 cited several examples to support her judgements and she evaluated grammar thoroughly and in more details than the descriptors required on the scale. She made 19 comments on evaluating grammar, for example: "Here the writer chose a wrong auxiliary".

The other German rater, GerR4, followed a similar route; however, she justified her judgements 23 times. After making the decision and announcing the score, she explained why she awarded that particular score, for example, "I would also like to mention in connection with this letter that the communicative goal has not been achieved, that's why it is a 0". She made a total of 16 comments on quality of script, saying, "Not very bad, it is good". This rater had the highest number of blank task sheets, she evaluated 36 scripts altogether, out of which 20 did not contain any or sufficient language. When she came across one or two empty task sheets and started the next one with some language on it, she said, "S/he also wrote very little, but the point is that at least s/he has written something".

6.6 Conclusion

The analysis of verbal protocols produced during the rating of written performances shed light on numerous features of rater behaviour. As regards the way raters arrived at the scores, I can conclude that it is not an easy task to exclude subjectivity during rating. Cohen (1994) mentions the influence of expectations on the rating process. In addition, the rating task in itself was significantly different from the everyday testing practice. The five raters who took part in the research did not have substantial experience in testing and marking a large number of scripts was completely new to them. In addition, they had little knowledge of the learners' background, which would have influenced their judgements. In some cases, they still attempted to make judgements considering surface features and they estimated learners' proficiency based on the scripts.

As the results show, the five raters mainly followed the procedure presented at the training. However, there were some differences in the assessment process. The reading pattern was similar, second reading occurred only if there was uncertainty in awarding the appropriate score; raters either compared the scripts to each other or changed their mind in connection with the score, so they reread the script to justify their second decision.

Findings of the research show that the rating process can be traced. The raters' moves were apparent, which can help to make rating more predictable and objective. What is more, raters could take appropriate measures, which they got familiar with during the training and thus felt more comfortable when assessing the scripts. It is also true that in some cases, which could not have been predicted before the actual marking took place, such as irrelevance of the scripts to the task or insufficient language, raters had to find a strategy for solving the problem.

The think-aloud protocol as research methodology turned out to be a useful means for gathering data on raters' thinking processes, but it needs further refinement. As a focus for a further study, the same scripts should be evaluated with more raters. Transcribing the protocols also needs more consideration; a great effort was needed in some cases to transcribe the audiotapes. The coding scheme developed should be more straightforward as far as the codes are concerned and tried out with other think-aloud protocols to see how it works with different data. Sometimes problems occurred with short, unclear utterances, one-word remarks, as wording of thoughts was not clear enough. When looking at individual utterances, the whole sequence should be considered, as not all of them are comprehensible without bearing the context they appear in mind.

Chapter 7

The Main Study: Processes and Outcomes in Rating L2 English Written Performance in a Pre-Service Tefl Course

Introduction

This chapter presents the research design, aims, research questions, participants and procedures of the main study. I attempt to shed light onto several factors influencing raters' rating behaviour in written performance assessment. Investigation into raters' decision-making processes using their verbalised thoughts can reveal what they focus on in rating, how they interpret the scripts, the rating criteria and how they employ the rating scale. Participants of the study are novice raters who are divided into competent and proficient groups according to their rating performance. I intend to investigate rating processes by comparing competent and proficient raters' behaviour. In order to observe raters' behaviour, I use think-aloud verbal protocol analysis as a methodology for collecting data on raters' thinking.

7.1 Background to Main Study

As a teacher trainer and a member of the British Council Examination Reform Project team I have been interested in and researched (Bukta, 2000; 2001; 2007; Bukta & Nikolov, 2002) L2 assessment, particularly written performance assessment for several years. I have been teaching an elective course on English language testing at the University of Szeged, Hungary with the intention of passing on the expertise I gained while working with the Project team. My experience shows that teacher trainees are interested in testing in EFL, especially because they can utilise their skills and knowledge in language testing as soon as they get to school to complete their teaching practice. In addition, they often reflect on the way testing is dealt with in different schools in Hungary and express their worries in connection with it. The issue of testing L2 performance has been in the centre of attention since preparations for a new school-leaving examination started in 1996. Although L2 education in Hungary has undergone substantial

changes over the last decades, not enough empirical research has been carried out to look into L2 written language assessment.

In order to observe raters' decision making processes, I conducted a pilot study, as presented in Chapter Six. The aim of the pilot study was to reveal what features raters attended to when evaluating written performance and I intended to elaborate on the means of data collection, which was verbal protocol analysis. The participants were five raters who provided think-aloud protocols while rating scripts written by Hungarian FL students. Findings showed that it was possible to trace raters' decision-making processes and they mainly followed the rating procedures presented at training. However, their rating patterns showed great variability and they often attended to different features. In addition, the study made it possible to pilot data collection procedures and ways of processing verbal data.

Findings of the pilot study generated the ideas for further research and provided a basis for deeper inquiry into raters' thinking. The main study was carried out to investigate how raters arrived at a decision while rating the same ten scripts. The main aim was to explore the way raters interpreted written samples and the rating scale, compare their decisions and trace their rating processes. Thus, the aims of this study are manifold: I would like to contribute to the research in written performance assessment in the Hungarian context. In addition, I would like to draw colleagues' attention to the importance of the role raters play in assessing writing skills and raise their awareness in dealing with language testing issues at both pre-service and in-service levels. Finally, think-aloud methodology as means of collecting verbal data can stimulate further research in the field.

7.2 Design of the Study

Complexity of written performance assessment means that in order to understand how raters make their decisions and what criteria they employ needs investigation from several aspects. First, the results of rating, namely the scores provide valuable information about performance and rater characteristics. To be able to relate raters' scores to the way they arrived at their decisions it is necessary to examine their thinking processes. Verbal protocol analysis is the method chosen for getting an insight into raters' thinking. The 37 raters were verbalising their thoughts as they were assessing the ten scripts and they had to produce an audio recording which they were asked to transcribe. Although mental processes are not directly observable, inferences can be made on raters' cognition on the basis of their thought articulations (Ericsson & Simon, 1993; Lumley, 2000).

7.3 Research Questions

Even though raters award scores to pieces of written performance using a rating scale after having been trained for the rating exercise, there are often differences between scores based on raters' decisions (Alderson et al., 1995; Bachman, 1990; Bachman & Palmer, 1996; McNamara, 1996; Weigle, 2002). Some raters agree with each other more, while some agree to a lesser extent. To reveal more about the nature of rating processes and get a better insight into some of the characteristics that play a role in raters' decision-making processes, raters can be grouped according to their rating performance (Weigle, 1999; Wolfe, 1997). All participants of the present study are novice raters as far as they rating experience is concerned. Still, their agreement in awarding scores shows different patterns, which allows them dividing into competent and proficient groups for investigating features of rating processes. Comparisons between the two rater groups are made to shed light on raters' rating behaviour. Thus, the following five broad research questions are posed:

1. What features characterise competent and proficient raters' rating patterns?
2. What criteria do competent and proficient raters focus on during rating?
3. How do competent and proficient raters interpret the four rating criteria?
4. How do competent and proficient raters interpret the scripts?
5. What is raters' feedback on the rating task?

7.4 Participants

Thirty-seven raters were involved in the rating exercise for this study; they assessed the same set of ten compositions and awarded scores on four rating criteria using an analytic rating scale. They were students from two university seminar groups; 32 female and 5 male English teacher trainee students (see Table 7.1 for details), who signed up for a course on testing in English language teaching. The course was held in two consecutive semesters in academic year of 2004/2005 at the University of Szeged, Hungary. Raters in the rest of the study are referred to by identification numbers, participants in the first group are referred to from R1 to R19, and in the second group from RR1 to RR18.

Table 7.1
Distribution of Participants in Two Seminar Groups

Group/Year	Rater identification	Gender		Total
		Female	Male	
1st group/2004	R1 – R19	14	5	19
2nd group/2005	RR1 – RR18	18	0	18
Total		32	5	37

All students participated actively in the seminars and handed in the rating assignment. They were Hungarian by nationality except for one male student, who was an ERASMUS scholarship student from Italy. The Hungarian full-time pre-service teacher trainees at the University of Szeged had received the same input in ELT methodology. The gender distribution in favour of female students (87% female and 13% male students) is characteristic of the English major teacher-trainee student population at this university. Some of the students had started their teaching practice parallel with the seminar course or had already had some teaching experience. They are considered novice raters as none of them had any experience in testing written performance before the rating exercise conducted for the present study, and had never taken part in any kind of rater training. As the participants' background is similar, the two seminar groups are treated as one sample and referred to as raters in the following sections of the study.

Raters assign scores to written performance. These scores provide information not only about candidates' language proficiency, but they can be used to examine agreement between raters. Inter-rater reliability is a measure indicating agreement between raters and it allows us to distinguish proficient raters from less proficient ones. Research into rater behaviour compares "experienced" and "less experienced" raters, where "experience" refers to the degree of expertise in rating. This expertise is referred to as "proficiency" in rating and indicates the degree of agreement among raters and not the experience in rating (Wolfe et al., 1998, p. 87). Elsewhere, such distinction of raters according to their expertise or rating proficiency is referred to as "good", "average" and "poor" (Green, 1998). Thus, considering the degree of agreement, raters can be grouped based on inter-rater agreement. Participants of this study are all novice raters with similar background features, so their grouping, as presented in Procedures, considering their agreement with the benchmarks. Benchmark is a standard against which measurement is conducted (Davies et al., 1999). The benchmarks in the present study are the researcher's scores, as I have substantial experience in rating. The awarded scores serve as a point of reference for further investigation.

7.5 Data Collection Instruments

7.5.1 The Scripts

The writing task was a guided composition in form of a letter of 150 words to a British friend (Bukta, Gróf, & Sulyok, 2005, p. 23, see Appendix 10.1). There were five content points, which the candidates had to elaborate on. The task corresponds to the examination specifications of the intermediate-level Hungarian school-leaving examination in EFL. The written component of the intermediate level school-leaving examination targets CEFR levels A2 and B1 (*Részletes vizsgakövetelmények [Detailed examination requirements],* 2003) and it consists of two tasks, one of which, Task A, is a transactional letter on a topic familiar to the students.

The compositions were written by Hungarian secondary-school students who were preparing for the school-leaving examination at the secondary school affiliated to the university in autumn 2004. The scripts were not produced in an operational setting, similarly to other studies inquiring into raters' behaviour (Lumley, 2000). They were written under exam preparation conditions: 24 students volunteered to try out the examination task in one of their regular English classes. The 24 compositions represented the pool of scripts out of which ten were selected for the study and the rest was used in the rater training sessions. The script selection procedure and rater training are introduced in the Procedures section below.

7.5.2 The Rating Scale

A six-point equal interval analytic scale was used (see Figure 7.1; a copy of the original can be found in Appendix 7.2) for rating. It is divided into four rating criteria to assess task achievement, vocabulary, grammar and organization. The principle of assessing written performance according to four criteria followed the way of designing analytic scales for rating writing tasks for the new school-leaving examination model. Written performance assessment was not included in all school-leaving test booklets before 2005, if they were, the assessment criteria were vague.

Points	Task Achievement	Vocabulary	Grammar	Organization
6	☐ Achieves communicative goal; the letter is for a friend ☐ All 5 content points covered	☐ Wide range of appropriate words and expressions	☐ Only one or two inaccuracies occur ☐ Structures correspond to task	☐ The layout fully corresponds the task ☐ There is clear logical link between all text levels
5				
4	☐ Communicative goal mostly achieved ☐ Almost all content points covered	☐ Good and appropriate range of words and expressions	☐ There are some inaccuracies but the whole text is comprehensible ☐ Some variety in structures	☐ The layout reminds of a letter ☐ There is some link at most text levels
3				
2	☐ Strain on the reader to comprehend ☐ 2 or 3 content points covered	☐ Basic words and expressions	☐ Basic mistakes hinder comprehension ☐ No variety in structures	☐ The layout is inappropriate ☐ There is no clear logical link between the elements of the text
1	☐ Communicate goal is difficult to comprehend ☐ 1 content point covered	☐ Very limited range of words	☐ Only part of text is comprehensible ☐ Most structures inaccurate	☐ The layout is messy ☐ Logic is missing between the different elements of the text
0	☐ Did not write anything or just some words ☐ Misunderstood the task			

Figure 7.1. Analytic rating scales for the letter-writing task

Consequently, rating scale development for the new examination had little tradition to rely on. As the new examination model was designed along the lines put forward in CEFR (2001), so a team of experts developed an analytic rating scale considering not only CEFR guidelines, but also scales used in respected international examinations (Szabó, Sulyok, & Alderson, 2000, pp. 62-65).

The design principles of the rating scale for the model examination include four aspects each reflecting the components of language ability which are measured. An analytic scale is one type of measurement instrument used in written performance assessment with the help of which, as discussed in Chapter Four, raters can give a detailed picture of various components of learners' writing

ability. Each of the components can be awarded with a separate score showing differences in performance in each aspect of ability (Alderson et al., 1995). In order for the raters to be able to distinguish ability differences in performance, the analytic scale should be straightforward and descriptors should be explicit and easy to use (Weigle, 2004).

The analytic scale for the present study was developed bearing in mind design principles established for the model school-leaving examination. The main principles were: the scale should reflect the measured writing ability; it should not contain too many descriptors; the descriptors should be straightforward and brief; there should be identifiable differences between levels of ability, and finally, raters should have a possibility for detailed distinction without awarding fraction scores. Last, the analytic scale should be tried out and raters should be trained to use it.

The analytic rating scale is broken down into the four rating criteria and seven band levels. Descriptors of each of the four criteria are grouped into bands, task achievement, grammar and organisation have two descriptors, and there is one descriptor for vocabulary. There are seven bands altogether to distinguish between levels of ability. Although it is a seven-band scale, only four bands contain descriptors and there are two so-called "empty" bands for making more detailed judgement possible. If a script exhausts only one of the descriptor criteria of a band and it is above the band descriptors below it, the score between the two bands can be given. This way more detailed evaluation is possible and fraction scores are avoided. The total score on an aspect is 6 points, and the total top score is 24 points in this analytic scale. There is a seventh band for task achievement, band 0, which has two descriptors: if the script does not contain sufficient language for evaluation or if it does not meet the task requirements, it is awarded 0 as a final score and no further assessment is needed.

The layout of the scale is transparent, the original is printed on an A4 sheet in landscape format; tabular arrangement aids orientation and the descriptors are allocated in bullet points (see Appendix 7.2). All descriptors are short, they have been thoroughly explained, and trialled on training sessions, as described below. The two descriptors of task achievement are to assess the degree of achieving communicative goal, that is whether the letter is for a friend; and how many out of the five content points are covered. Vocabulary is assessed according to the range and degree of appropriacy of words used by students. Grammar descriptors comprise language accuracy assessment and variety of structures. The writing task, as defined in the prompt is to write a letter to a friend that is why the first descriptor of the aspect of organisation is assessing the layout, how much the text follows the formal requirements of a letter. The second descriptor in this aspect is assessing coherence of the text.

7.5.3 The Rating Task Assignment Package

The rating task, during which the data were collected for the study, was meant to be a home assignment for students to be completed on their own during the examination session period and handed in, as agreed, by the end of the examination session. As students had to work on their own, data collection instruments had to be designed carefully. Before going into the details of rater training, the data collection instruments are introduced, which were collected in a rating pack. The rating pack was an envelope labelled with the students' name and handed to them at the end of the training sessions.

The Letter to the Student

Although the rating task was carefully explained and both the rating procedure and think-aloud protocol production had been piloted during rater training sessions, the rating pack contained a checklist for the students in a form of a letter as a reminder of what they were expected to do (see Appendix 7.3). The letter contained the most important instructions and information students had to bear in mind before, during and after rating. First, they had to check how the ten scripts were labelled, check the quality of audio recording, and look at the writing task and the rating scale. It was also suggested that they should revise seminar notes to see what was discussed in the training sessions. I provided my contact address in case a problem occurred. The instructions during the rating task included a reminder for reading the scripts at least twice, as practised in rater training, and technical reminders to announce the script identification, to enter scores in the score sheet and to keep talking. After rating, students were required to produce the transcript of the protocols and to fill in the feedback sheet. Finally, the list of items in the envelope was there for students to check that they would hand in all items requested.

The Scripts

The rating pack contained the ten scripts (see 7.5.1 for detailed description) which had been selected carefully, as presented in Procedures below. The scripts were numbered from 1 to 10 and are identified in the study as script N1, script N2, etc. The ten scripts were Xerox copies of the original scripts of handwritten texts (see Appendix 7.4 for a copy of each). Different features of handwritten scripts influence raters' impression of the text quality (Shaw, 2003). Although scripts are often typed for research and training purposes, typed versions of original handwritten scripts may influence rating, as the findings show in case of comparison of computer-assisted written performance assessment and paper-based written performance assessment (Lee, 2004).

The Rating Scale

Although the same rating scale, as introduced above, had been used in the training sessions, a copy of it was included in the training pack in case the other one was lost.

The Score Sheet

The training pack contained a score sheet: a grid with the script identification numbers (N1 to N10) in columns and the four criteria as they appeared in the rating scale (see Appendix 7.5).

The Feedback Sheet

The feedback sheet (see Appendix 7.6) invited the raters to comment on the seminar course on testing in EFL, on the training they had for the rating task, and on the rating task itself. The rationale for not presenting a structured questionnaire was to avoid any influence by posing specific questions. Data collected this way may not be easy to analyse: still, the aim was to get feedback on how students perceived the seminar course, the input they got in language testing and the rating task.

7.6 Data Collection Procedures

To answer the research questions both quantitative and qualitative data have been collected:

1. quantitative data are the scores awarded by 37 raters and the benchmark scores on four criteria of ten scripts written by secondary-school students
2. qualitative data were collected using think-aloud verbal protocols: the 37 raters recorded their verbalised thoughts on an audio tape and transcribed the protocols; therefore, a total of 370 protocols were produced

Data were collected in the examination session of the 1st and 2nd semesters of 2004/2005 academic year after teacher trainees had completed a one-semester seminar course in testing in ELT. The course (the course description can be found in Appendix 7.7) served as rater training for the rating exercise of the present study. The following discussion focuses on script selection and grouping raters. Rater training is dealt with in detail to explain how students were trained for the rating task. Then, the way verbal data were processed was introduced. If verbal report data are collected, data collection should be described in as much detail as possible (Ericsson & Simon, 2003), that is why each phase of data collection is detailed below.

7.6.1 Script Characteristics

The writing task used in the present research (see 7.5.1 for detailed description) was a guided composition: students were to write a letter to a friend and include five content points. The four categories of the analytic rating scale were considered as items. Thus, Cronbach α was calculated for the dataset of 37 raters on each of the four criteria in the analytic scale and for all four criteria considering one aspect as an item. Reliability coefficients show that the writing task measured the targeted component of language ability consistently (see Table 7.2) at an acceptable level: the reliability estimates are between .70 and .91.

Table 7.2
Reliability of Scripts (Cronbach α)

Criterion	Cronbach α
Task achievement	.76
Vocabulary	.74
Grammar	.70
Organisation	.81
Total score	.79
TA+V+Gr+O	.91

To look at the characteristics of the ability that the task intended to measure, cluster analysis of the four rating criteria was carried out which revealed that linguistic features (vocabulary, grammar and organisation), are separate from task achievement, the content measure (see Figure 7.2).

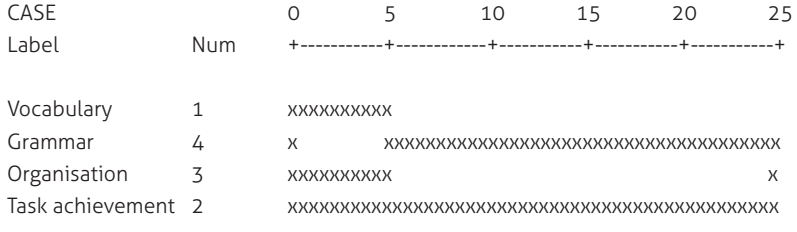


Figure 7.2: Cluster analysis for the four rating criteria

7.6.2 Script Selection

The ten compositions for the study were selected from a pool of 24 compositions written by secondary school students. The rearcher who has a substantial experience in rating first rated all 24 scripts and the awarded scores constitute the benchmarks, which represent points of reference for the study. The benchmarks were then used for script selection and profiling raters; thus, ten scripts were selected for the research and the remaining fourteen scripts surved rater training purposes.

Table 7.3
Benchmarks in Points and Percentages of Total Scores for the Ten Scripts

Script number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Total score of 24 points	13	9	10	10	19	24	15	17	10	18
Percentages (%)	54	38	42	42	80	100	62	70	42	75

The intention was to provide a balanced sample of the top, medium and weak performances together with some that seemed problematic because of their content, length or layout. When the ten scripts were selected, they were randomly numbered from 1 to 10 to make sure there was no hint to text quality in sequencing (see Table 7.3 for benchmark scores).

7.6.3 Training Raters

The rater training for the rating exercise was built into the one-semester elective university seminar course. At the beginning of the Testing in ELT elective course, students received information about the research and the rationale.

The course on testing is one of the elective courses offered in the last phase of training for an English teaching degree at the University of Szeged, Hungary. The seminar comprises a double session each week lasting 90 minutes for 13 weeks of the semester. The aim of the course is to familiarise teacher trainees with the main trends of theory of L2 testing and to provide some practice in language assessment (see Appendix 7.7 for course description). First, awareness raising tasks demonstrate the relationship between teaching and testing, and highlight the important role assessment plays in language instruction. Then, the basic theoretical issues of language ability and its assessment are discussed. After the theoretical input students have an opportunity to look into practical issues of testing language ability components including written performance assessment.

The difference between the regular schedule of the seminar course and the one with the rating task for the present study was that students received more detailed input on written performance assessment and had rater training sessions. In addition, the course requirements included the rating assignment instead of a more traditional way of assessing students (the course requirement usually is a short summary of reading the relevant literature). The first part of the course followed the syllabus and the last four seminar sessions were allocated to preparation for data collection for the present study.

The sessions devoted to discussions of written performance assessment started with the explanation of the goal of the present research and the assignment. After an introduction on written performance assessment in general, the focus shifted to the rating task. Students went through three rater training sessions, in which they studied the writing task and the rating scale, and they had an opportunity to practise rating. They could also see in practice how inter-rater and intra-rater reliability work. In addition, think-aloud protocol as methodology for collecting verbal data was explained and tried out.

The training of raters followed the scenario of the training procedure described in Chapter Four and suggested by Alderson, Clapham and Wall (1995, pp. 105-127). The goal of the training was to introduce the notion of subjective assessment and to practise rating using an analytic rating scale. It was also important to standardise rating for the rating exercise and to familiarise students with the analytic scale and the task.

First, in the training sessions raters studied the writing task, discussed the requirements concerning content as described in the five content points of the rubrics and the expected language features. Then, the introduction of the analytic rating scale and the scale descriptors followed. I chose five benchmarked scripts from the pool of 14 scripts that remained after the first selection of ten for the rating task. After rating these five scripts, students got detailed justifications and they had to compare their scores to the benchmarks and reach consensus. There were three further scripts that students had to rate. Although elimination of all emerging problems during rating is not possible, training should provide sufficient practice for raters to arrive at informative judgements (McNamara, 1996, p. 127). Students had an opportunity to rate eight previously benchmarked scripts and took part in detailed discussion of the scores awarded and they could compare their scores to the benchmarks. Students practised vocalising their thoughts in pairs and they commented on their own and each other's performance on thinking aloud. The activities with which they practised think-aloud were similar to the rating tasks, students had to rate scripts verbalising their thoughts. Finally, students received instructions for the rating task and got the rating packages.

7.6.4 The Rating Task

Data for think-aloud verbal protocol reports can be collected with the researcher present or in her absence. Both have their advantages and disadvantages: if the researcher is present, she can maintain the verbalisation of thoughts by prompting. However, this can be a disadvantage, as the researcher may influence the thinking process (Ericsson & Simon, 1993; Green, 1998). For the present study, the researcher's presence during rating did not seem feasible, so the instructions and checking the technical background were of paramount importance. The pilot study on tracing raters' cognition, as discussed in Chapter Six, showed that valuable data can get lost unless all aspects of data collection are considered. Experience with think-aloud protocol data shows that if the researcher is not present during verbalisation, it is advisable to ask for both the audio recording and the transcripts of data. Moreover, participants of the present study were teacher trainees and the transcription exercise raised their awareness in rating written performance and they could also experience how research is carried out.

Students were motivated to complete the rating assignment, as they were graded and awarded credit for the seminar course on fulfilling the assignment in good quality. Deadline for submission of the assignment was discussed; they were aware of the fact that the rating task was time-consuming and were asked to devote as much time as needed after they had taken all their examinations for the semester. All 37 rating packs were returned with completed rating task, so no data got lost. The following stage was data preparation which meant processing the transcripts for further analysis.

7.7 Processing Verbal Data

The research type defines whether all data are transcribed or only part, and the personnel for transcribing can be identified (Ericsson & Simon, 1993). Raters in this study had to audio record all their talk and provide a transcript for it on the computer disc provided. Students were asked to type everything they said without making any alterations, so that the typed protocols were the written form of what they actually said (see Appendix 7.8 for a sample transcript). Thus, data processing started with comparing a randomly chosen sample of 74 (20%) protocols of audio recordings and their transcripts to get evidence that they matched. Students were not instructed to transcribe data according to transcribing conventions (Green, 1998), so the transcripts did not have time indicators. As the first step of data processing, I changed the typeset to distinguish the different text units according to the research aims, as described below.

7.7.1 Transcript Segmentation

Processing verbal data is an extremely time-consuming exercise. The framework for verbal data processing includes several stages that should be observed in order to make the research valid and reliable (Ericsson & Simon, 1993, p. 312). Transcript segmentation and coding scheme development are arbitrary processes, posing several dilemmas for the researcher, and the principles are set according to the research aims (Green, 1997). Thus, the first decision to be made is how to segment the text. Dividing the text into text units (TUs) allows the researcher to focus on the aims of research (Lumley, 2000, p. 125).

Bearing in mind transcript segmentation principles put forward in the literature and the experience gained in a similar study in which think-aloud protocol was used as data collection methodology, as described in the pilot study in Chapter Six, transcript segmentation was carried out. The two main principles for segmentation were: to make the transcripts analysable to answer the research questions and to make them transparent.

As far as research questions are considered, the transcripts were segmented into meaningful text units: (TUs) which referred to

1. comments on the rating process,
2. scoring criteria,
3. what raters attended to: the script, the scale or came up with their new criteria.

Transcript segmentation raised several dilemmas: the first one was completeness of comments. Transcripts are recordings of thought processes and raters focused on the rating task not on completeness of utterances. That is why segmentation followed the principle of thought content not linguistic features: most comments are incomplete utterances, as illustrated in Excerpt 7.1: TU3 is error identification; in TU4 the rater identifies lack of detail; TU6 is rating category identification and in TU7 R9 chose the score.

Excerpt 7.1: An example of incomplete utterances

R9 Script N9

TU	Rater talk
1	Now let's move on to script N9
2	I'm very honest that you could ... leaving ...What I hate I had ... sovereign
3	also a problem
4	not talks about ... yes! Inviting programme ...
5	Looking forward your arriving
6	task achievement
7	4

The decision regarding these deficiencies was to segment the protocols according to the identified focus and not to correct raters' mistakes following principles in earlier research (Lumley, 2000; 2002).

The distinction between script reading and providing an example posed a further dilemma in segmenting. Sometimes raters cited examples from scripts which were occasionally lists of words, whereas in other cases they were reading scripts out. The dilemma of what was considered an example and what an extensive text was solved by segmenting individual words as examples, and more than one word as script reading as in Excerpt 7.2.

Excerpt 7.2: Two types of reading script: reading examples and reading extensive text segments

RR9 Script N4

TU	rater talk
21	and *a shoes*
22	*a chocolates*
23	*lots of present*

RR9 Script N1

TU	Rater talk
9	like *at the end*
10	*This will be a good party* and ...
11	about the train... *the train started at the same time ...*

Transparency means that it was important to create a layout in which not only stages of rating processes were apparent, but also what the different comments referred to. As the example in Excerpt 7.3 shows, the comments referring to rating criterion nomination (TU4 and TU6) and score selection (TU5) are separate text units to aid following rater's thinking.

Excerpt 7.3: An example of transcript segmentation for transparency of transcripts

RR9 Script N1

TU	Rater talk
1	The content points...he writes about all the five content points.
2	It's understandable and it's
3	...the communicative goal is achieved, it's a letter for a friend.
4	I think task achievement
5	is five points.
6	Vocabulary.

Following the principle of layout transparency in transcripts resulted in employing different typeface for distinguishing different TU types. The following typeface was used:

1. normal print for raters' own talk,
2. boldface when they were reading the rating scale,
3. boldface underlined when they were reading the writing task instructions,
4. italics when they were reading the script.

To sum up, transcripts were segmented into TUs based on content; TUs were numbered and different typeface was used to distinguish them. Raters were referred to as R1 to R19 and RR1 to RR18, respectively, according to the group they belonged to (see 7.4 above). RRR indicates the benchmarks (my scores). Scripts are identified by numbers, for example: Script N1. These principles are illustrated in Excerpt 7. 4.

Excerpt 7.4: Examples for typeface use, rater, and script identification in transcripts

RR15 Script N1

TU	Rater talk
1	The student **achieves communicative goal, the letter is for a friend,**
2	**and all five content points are covered,**
3	although I think that she mentions some irrelevant things in her letter,
4	so I give her a five for that.

RR3 Script N2

TU	Rater talk
11	the **invite your friend for next holiday**
12	is only saying *Can you come?*
13	This seems strange to me this way and all sentences after *Can you come* there is no link between them

As the examples in Excerpt 7.4 show, the first and second segments in RR15's rating script N1 protocol are reading the first and second descriptors of the scale for task achievement, the third text unit is the rater's remark on the relevance of some parts of the text, in TU 4 the rater chooses a score, and, as the fifth segment shows, goes on and announces the next scoring category. As the other example shows, when RR3 is rating script N2, she is reading from the writing task prompt (TU11), then she is reading from the script (TU12), and evaluates that part of the text in her own words (TU13). When all protocols had been segmented as described in this part, protocol coding was carried out.

7.7.2 Coding Scheme Production

The next stage in verbal data processing involves identification of each TU and development of a coding scheme to make further analysis possible. Research based on verbal protocol analysis has one feature in common: all researchers agree that there is no uniform coding scheme and a specific one should be developed for each study (Ericsson & Simon, 1993; Green, 1997). Each one is conducted with different research aims and questions, using different materials and involving different participants. Although some categories can be decided a priori, the coding scheme develops as the data are being analysed. As far as the required details are concerned, "in coding, the categories should always be as narrow as possible" (Mackey & Gass, 2005, p. 230). However, as findings of the pilot study in Chapter Six show, too many categories are difficult to handle. That is why the maximum of 50 categories seems feasible.

Protocol segmentation had already raised some dilemmas and there was a further one that emerged in coding the TUs. This dilemma related to the language of the transcripts, as the rating task did not pose restriction on language use, raters produced three types of protocols: in English, in Hungarian and there were protocols using both languages. The language of the transcripts was not changed for segmentation and for coding: however, if raters' talk is cited in the study, it always appears in English, as I translated rater talk not in English. As far as linguistic accuracy of rater talk is concerned, rater talk has not been altered or corrected.

The coding scheme was set up bearing in mind findings of the pilot study and earlier research into rating processes. Thus, initially three main stages of rating processes were identified that were used as a starting point of coding scheme development for the present study. The three main behaviour types identified by Lumley (2000, p. 131) and Cumming et al. (2002) are:

1. management behaviours,
2. reading behaviours,
3. and rating behaviours.

Management behaviours were divided into two types: script or criterion identification (code 1) and remarks on the rating process (codes Da, Db, Dc and Dd), as shown in Figure 7.3.

The rationale for separating script and criterion identification is to aid data processing, as a separate code for identification shows change in raters' focus. Two examples in Excerpt 7.5 show how scripts and criteria were identified and how raters commented on the rating process itself. RR15 identified the rating criterion she was going to deal with (TU14) for script N5; the same rater added up scores when dealing with script N1 at the end of rating and announced the result (TU17).

Code	Comment
1	Identifies a script/criterion

Code	Focus	Code	Comment
D	Remarks on the rating process	Da	Repeats score
		DB	Adds up scores
		Dc	Unclear reasoning
		Dd	Comments on the rating process

Figure 7.3. Codes for two types of management behaviour

Excerpt 7.5: Two comments illustrate types of management behaviours

RR15 Script N5

TU	Rater talk
14	and organization

RR15 Script N1

TU	Rater talk
17	So the total is sixteen for this paper.

Research is directed towards identifying rating processes and raters' interpretation of the scale, so rating behaviours are related either to the scale categories or to the raters' own rating focus. In case of holistic marking, composition elements raters focus on can be identified (Milanovic et al., 1996) and data are coded according to what raters attend to. An analytic scale, as described in Chapter Four, consists of rating criteria to direct raters' attention towards text characteristics (Weigle, 2002). The rating scale for the present study, as introduced in 7.5.2, consists of four rating criteria: task achievement, vocabulary, grammar and organisation according to which scores can be allocated. As far as comments on the four rating criteria are concerned, the four broad category types were further divided and each of them comprise the same eleven comments. The reason for establishing equal categories for the four rating criteria was to make comparison of raters' attention to scale categories possible. An example of rating categories according to scale descriptors for task achievement is shown in Figure 7.4.

Code	Rating aspect	Code	Comment type
A1	Task achievement		
		A1a	Compares text to scale descriptor 1: communicative goal
		A1b	Compares text to scale descriptor 2: content points
		A1c	Evaluates aspect in own words
		A1d	Adds own criterion
		A1e	Chooses score
		A1f	Adds reason why that score
		A1g	Revises decision
		A1h	Identifies error
		A1i	Refers to lack of detail
		A1j	Changes focus/switches to different criterion
		A1k	Finalises the score

Figure 7.4. Codes for rating task achievement in the coding scheme

To illustrate how the codes were assigned to TUs, a fragment of a coded protocol is given in Excerpt 7.6.

Excerpt 7.6: An illustration of assigning codes to text units

R19 Script N2

TU	Rater talk	
1	Task achievement,	1
2	not all the points are covered,	A1b
3	it lacks in the 1st one, when he had to thank his friend...	A1i
4	he also lacks in the programmes he is planning,	A1i
5	furthermore the other points are developed in a simplistic way...	A1b
6	it isn't always easy to comprehend what he wrote,	Ce
7	so... 2 for task achievement.	A1e

As the example shows, all behaviour types related to one of the four rating criteria in the rating scale are included in one rating category.

Reading behaviours occurred when raters were reading the scale, the script, summarised the script content, cited one word as an example, or were reading the rubrics. This taxonomy is somewhat different from reading behaviour discussed in earlier studies (Cumming et al., 1997; Lumley, 2000; 2002), in which reading behaviours are those comments that are strictly related to reading scripts. Reading behaviour category in this study includes any occurrence of reading during the rating process: reading scale descriptors, script text or task rubrics, example or text extract. Raters' summary of the scripts or parts of scripts is also

added here. A distinction was made between giving an example and reading parts of a text. This principle emerged in protocol segmentation, as discussed above, so a distinction was established between providing an example and reading extensive text: a TU was considered as an example if it consisted of one word, any TU longer than one word was considered as reading the text. The rationale for creating separate categories for reading the scripts was based on an observation that raters very often provided examples to justify their decisions and they used reading parts of the scripts as a rating strategy. The dilemma of which reading type constitutes an example and which one belongs to a rating strategy was solved by considering the number of words that raters read out.

Examples in Excerpt 7.7 illustrate each reading behaviour comment type: R4 is reading the scale descriptor from the rating scale (TU31), the same rater is reading part of the script (TU7), R19 summarises script content (TU7), R16 is giving an example (TU12) and R16 is reading a content point from the rubric (TU3).

Excerpt 7.7: Five examples of reading behaviour comments

R4 Script N10

TU	Rater talk
31	**The layout reminds of a letter**

R4 Script N9

TU	Rater Talk
7	...starting with *I am very honest*

R19 Script N7

TU	Rater talk
7	but he just drops a few lines ... about a party next week and how could a friend come from England for next Friday.

R16 Script N12

TU	Rater talk
12	for example *typical*

R16 Script N3

TU	Rater talk
3	but when he talks about the programmes

These reading behaviour comments were classified according to the criteria they related to, as there were four rating criteria: task achievement, vocabulary

grammar and organisation, reading behaviour comments were coded accordingly. The example (see Figure 7.5) for reading comments for rating task achievement shows the five reading comment types identified.

Code	Rating aspect	Code	Comment type
B1	Task achievement		
		B1a	Scale
		B1b	Script: more words
		B1c	Summarises script
		B1d	Example: one word
		B1e	Rubric

Figure 7.5. Codes for reading behaviour comments on rating task achievement

There was another type of rating behaviour that could not be related directly to any of the rating criteria, so the label for these comment types is "other comments" and contains 13 comment types which represent raters' own focus, as shown in Figure 7.6.

Code	Focus	Code	Comment type
C	Other comments (own focus)		
		Ca	Reflects on length
		Cb	Reflects on eligibility, tidiness
		Cc	Reflects on quality of script
		Cd	Comments on overall impression
		Ce	Comments on comprehensibility
		Cf	Comments on observation of student proficiency
		Cg	Corrects error
		Ch	Reflects own feeling
		Ci	Reflects on relevance of content
		Cj	Meditates on student intention
		Ck	Suggests solution
		Cl	Compares to other script/student/score
		Cm	Expresses uncertainty

Figure 7.6. The "other comments" section of the coding scheme

These comments were global comments on different aspects of script quality, including surface features such as length, layout and comments on content including comprehensibility and overall impression. There were several comments reflecting raters' evaluation of candidates' situation, such as their proficiency or their assumed intentions. Raters' reference to text relevance, their

feelings generated by the texts, comparisons with other scripts or scores were included here and their reflections of uncertainty.

The coding scheme was developed gradually and several modifications had been made before the final draft was ready for use. Modifications were the results of several re-readings and re-scoring of some of the protocols. Thus, the coding scheme (see Appendix 7.9 for a copy) is divided into five broad categories and labelled as follows:

- A number, "1", was used to label script or criterion identification to make transcripts transparent and easier to analyse.
- Capital letter "A" for rating behaviours was divided further into four according to rating criteria: A1 for task achievement, A2 for vocabulary, A3 for grammar and A4 for organisation. Each of the four subcategories is further broken down into eleven subcategories marked from "a" to "k". For example, "A1h" means that rater identifies an error when rating task achievement.
- Capital letter "B" is a label indicating the different reading behaviours and is divided into four subcategories according to the four rating criteria: B1 for task achievement, B2 for vocabulary, B3 for grammar and B4 for organisation. Each of the four subcategories is further broken down into five subcategories marked from "a" to "m". For example, "B3c" means rater summarises script content when rating the aspect of grammar.
- Capital letter "C" is for labelling comments of raters' own focus and was marked from "a" to "m", for example, "Ce" indicates when raters commented on comprehensibility.
- Capital letter "D" is the indication of those comments that referred to the rating process and were marked from "a" to "d", for example, "Da" was used to indicate that rater repeated the score.

Thus, the coding scheme contained 82 different codes grouped according to raters' focus and an attempt was made to produce a coding scheme that could be handled easily both for coding and analysis (see Appendix 7.10 for a sample coded transcript).

All transcripts were segmented and coded during the 2004/2005 academic year. Finally, 88 protocols, that is 24% of the total of 370 scripts were re-scored by a second reader who, as it is suggested (Green, 2000, pp. 63-68), was not involved in the research. The second reader was a non-researcher rater (a second-year English major teacher trainee student) who volunteered to rate a randomly chosen sample of protocols. The result of inter-rater reliability is encouraging; the agreement is 85% between the codes I assigned and the second reader.

7.8 Rater Characteristics: Grouping Raters

The raters' role in written performance assessment is very important, "differences of opinion may indicate that scorers think differently about the features of the essay on which scores are based and about procedures used to read and evaluate the essay" (Wolfe, Kao, & Ranney, 1998, p. 466). Their characteristics can reveal their rating processes and what they attend to when making scoring decisions. Raters' characteristics can relate to their experience in scoring and a distinction can be made between experienced and inexperienced raters (Weigle, 1999), or by looking at inter-rater agreement three groups can be formed depending on level of agreement (Wolfe, 1997; Wolfe et al., 1998).

The 37 raters' proficiency in rating was compared to the benchmarks (see Appendix 7.11 and 7.12 for details). Five comparisons were made to allocate raters into groups: agreement on each of the four rating criteria and agreement with the total score of each rater, which was calculated using Pearson correlation estimates. Correlation coefficients in Table 7.4 show different levels of agreement, thus, two groups of raters were formed: "competent" raters correlations were significant on three or fewer aspects, whereas "proficient" raters correlations were significant on four or five aspects with the benchmarks. Correlation estimates of all 37 raters with the benchmarks on each of the four scoring categories and the total score were calculated. The correlation coefficients of those raters who were assigned to the competent group are shadowed, and proficient raters' correlation coefficients are not. There are 22 (59% of the total of 37) raters in the competent group and 15 (41% of the total of 37) raters in the proficient one. From now on, raters of the present study are referred to as "proficient" and "competent" raters.

Table 7.4
Correlations between the Benchmarks and 37 Raters' Scores on Four Criteria and Total Score

RRR Pearson correlation	Task achievement	Vocabulary	Grammar	Organisation	Total
R1	.788**	.575	.505	.728*	.759*
R2	.881**	.654*	.859**	.851**	.911**
R3	.744*	.356	.793**	.782**	.709*
R4	.455	.563	.583	.735*	.722*
R5	.658*	.551	.529	.705*	.649*
R6	.615	.709*	.709*	.817**	.784**
R7	.228	-.079	.-709	.366	.115
R8	.851**	.712*	.692*	.739*	.797**
R9	.534	.315	.776**	.783**	.692*
R10	.820**	.335	.581	.791**	.710*

RRR Pearson correlation	Task achievement	Vocabulary	Grammar	Organisation	Total
R11	.649*	.153	.441	.748*	.605
R12	.538	.272	.749*	.773**	.779**
R13	.602	.543	.505	.813**	.741*
R14	.858**	.599	.732*	.835**	.838**
R15	.561	.358	-.053	.735*	.459
R16	.796**	.581	.599	.800**	.800**
R17	.573	.591	.697*	.528	.638*
R18	.266	.526	.670	.795**	.605
R19	.823**	.565	.467	.800**	.741*
RR1	.615	.742*	.705*	.691*	.819**
RR2	.534	.607	.619	.718*	.705*
RR3	.841**	.455	.867**	.786**	.804**
RR4	.871**	.737*	.689*	.922**	.921**
RR5	.423	.484	.791**	.839**	.819**
RR6	.573	.115	.446	.499	.580
RR7	.763*	.518	.822	.583	.776**
RR8	.678*	.806**	.673*	.819**	.804**
RR9	.788**	.402	.541	.533	.638*
RR10	.481	.366	.299	.603	.530
RR11	.635*	.477	.772**	.743*	.797**
RR12	.657*	.948**	.688*	.792**	.840**
RR13	.803**	.579	.712*	.806**	.773**
RR14	.565	.610	.705*	.532	.663*
RR15	.485	.711*	.651*	.700*	.723*
RR16	.932**	.413	.673*	.771**	.641*
RR17	.742*	.424	.524	.726*	.690*
RR18	.805**	.790**	.849**	.879**	.905**

*** Correlation is significant at the 0.01 level (2-tailed)*

** Correlation is significant at the 0.05 level (2-tailed)*

a Listwise N = 10

7.9 Summary

This chapter introduced the research carried out on raters' rating processes in rating written performance. Research into raters' rating processes focuses on identifying the stages of rating processes and what criteria raters pay attention to and, finally, how they arrive at a score. Frameworks of rating processes demonstrate the complexity of raters' behaviour and the wide range of criteria

that raters attend to in rating. Raters' characteristics play a central role in written performance assessment, their rating behaviour needs attention.

The present study attempts to contribute to the research conducted so far and to reveal more about the way written performance is assessed. First, the research aims at investigating raters' rating patterns by making several observations on their rating processes. Second, raters' focus is considered to find out about the criteria they attend to in rating. Third, raters' interpretation of the four rating criteria is looked into. Fourth, the differences in script interpretations are examined through following competent and proficient raters' rating of two scripts. Fifth, raters' perception of the rating task is discussed to conclude how would-be teachers of English see language-testing issues in general and written performance assessment in particular.

After posing the research questions, I introduced the participants, 37 students who were in their last phase of ELT training course and research instruments. The data were collected when raters were rating the same ten scripts with an analytic scale and they were verbalising their thoughts. The rating pack included a further data collection instrument, which was a score sheet to record the scores. Procedures comprised rater training for rating the scripts and training for the rating task including think-aloud data collection methodology. Then, the verbal data processing procedures were detailed to show how the raw dataset of verbal protocols was transformed into analysable data. Findings of the pilot study discussed in Chapter Six were utilised in segmenting verbal protocols and coding them in a manageable way. In order to have a clearer picture of rater characteristics, raters of the study were divided into a competent and a proficient rater group according to their rating performance. Finally, the principle for dividing the raters into two groups is explained. The following chapters focus on the research questions and discuss them one by one.

Chapter 8

Features of Raters' Rating Patterns

Introduction

The aim of the chapter is to present the results by answering the first research question. Raters go through the rating process according to certain behaviour patterns employing different strategies to arrive at a decision. I attempt to answer the first research question by looking into competent and proficient raters' behaviour during rating.

1. What features characterise competent and proficient raters' rating patterns?

As the description of the participants of this study in Chapter Seven shows, raters are all novices as far as experience in rating is concerned. They were assigned into competent and proficient groups based on their agreement with the benchmarks. Comparisons of the two groups are made to reveal more about the rating processes and to look at the emerging rating patterns. That is why raters throughout the study are referred to as competent and proficient and not novice raters.

Regarding the characteristics of competent and proficient raters' rating patterns, first I will look at gender distribution. Then, I observe features of script sequencing, which due to the low number of scripts and the narrow research focus can only be tentative. The intention is to reveal whether raters deviated from the order the scripts were numbered in the rating packages. Then, the language of the protocols is looked at to see which language raters used when rating the scripts. Raters' verbal protocol lengths are compared to investigate whether there is any relationship between the order they rated the scripts and the length of the protocols.

Next, raters' sequences of rating behaviours are compared to see what patterns emerge. An analysis of emerging patterns is carried out to answer the following specific research question: What patterns do raters follow when rating written performance?

Raters' identified behaviours are grouped into several categories, each of these categories include reading, rating and management strategies, which have further characteristics (Cumming et al., 2002; Lumley, 2000; Wolfe, 1997). The discussion that follows attempts to shed light on them to reveal features of the rating processes. By comparing the processes that competent raters follow

with those of proficient raters' thinking the intention is to see whether they go through the rating process in a similar way.

8.1 Raters' Gender Distribution

There was no specific participant selection principle for the study: all students who signed up for the elective course in testing in ELT were informed about the research and volunteered to take part in it. Gender distribution approximately reflects the gender distribution of teacher trainees majoring in ELT at the university. Unfortunately, there is no exact data available, but my teaching experience of more than 15 years at the university confirms that female students represent the majority of teacher trainees. That is why the ratio of 86 percent female (32 students) and 14 percent male (5 students) students is not surprising, as Table 8.1 shows. The two groups' gender distribution is similar, which was unintentional as far as raters' selection into the two groups according to gender was concerned. Raters' grouping was done, as discussed below, according to their agreement with the benchmarks and no other characteristic was taken into account. Thus, the competent group's gender distribution is: 86 percent (19 students) female and 14 percent (3 students) male raters. The ratio of female and male participants in the proficient group is: 87 percent (13 students) female and 13 percent (2 students) raters.

Table 8.1
Raters' Gender Distribution (percentages in parentheses)

Gender	Female	Male	Total
Competent raters	19 (86%)	3 (14%)	22
Proficient raters	13 (87%)	2 (13%)	15
Total	32 (86%)	5 (14%)	37

As the total number of participants in the study is low, the conclusions regarding gender distribution can be tentative. We can conclude that there is hardly any difference between the gender distribution of competent and proficient raters: the ratio of female and male raters is similar in the two groups.

8.2 Language of the Verbal Protocols

The rating environment was Hungarian, students were all of Hungarian nationality except one Italian student, so restriction regarding language use was

not made. Earlier research in rater behaviour investigation implementing think-aloud protocol was conducted in English and the issue on language use was not dealt with. As mentioned earlier, raters were instructed to use the language they felt most comfortable with, except for the Italian student, who was asked to use the English language in rating. Three different types of protocols were produced regarding language use: protocols in L1 Hungarian, L2 English, and in some cases raters used a combination of L1 and L2 (see Table 8.2).

There were more proficient raters (10 raters, 67%) who used both L1 and L2 than competent raters (4 raters, 26%), whereas half of the competent raters (11 raters) used only the English language while rating the scripts.

Table 8.2
Language Use in Verbal Protocols (percentage in parentheses)

Language	Hungarian (L1)	English (L2)	L1&L2	Total
Competent raters	2 (9%)	11 (50%)	9 (41%)	22
Proficient raters	1 (7%)	4 (26%)	10 (67%)	15
Total	3 (8%)	15 (41%)	19 (51%)	37

The Hungarian language was used by a total of three raters (8% of all raters), two (9% of all competent raters) of them were from the competent group and one (7% of all proficient raters) from the proficient group.

To sum up language use in rating, we can see that raters mostly used a combination of L1 and L2 in rating and proficient raters' talked in L1 and L2 more than competent raters, who spoke more in L2 during the rating task.

8.3 Sequencing the Scripts for Rating

Each rater received the same set of ten scripts in the rating pack which were numbered 1 to 10 and arranged in the same order with no instruction regarding what sequence to follow in rating. The principle of script arrangement was to offer a random order as far as quality is concerned. This principle is used in rater training sessions to ensure that raters are not influenced by script sequence (Alderson et al., 1995, pp. 105-127; Weigle, 2002, pp. 130-131).

Some raters deviated from the given script order and rated the scripts starting with the last. There were two raters, R4 and R15 in the competent group who started rating with script number 10, and three in the proficient group, who deviated from the order in the same way. In addition, R2, who belongs to the proficient group, rated the scripts in random order. She did not explain why she mixed them, and rated the scripts as follows: N5, N8, N1, N9, N6, N4, N3, N7, N2,

and N10. The other two raters from the proficient group, R3 and R8 started rating with the last script in the pack.

Although the low number of scripts and the focus of the study allow tentative conclusions as far as sequencing is concerned, some observation can be made. None of the 37 raters made any comment on the sequence of the scripts, even if they compared them while rating. There is no evidence why five raters (14% of the total) deviated from the sequence in which the scripts were presented. The tentative observation is that changing the order of scripts for rating influenced the raters' rating proficiency to some extent, as according to their agreement with the benchmarks two competent raters (9% of all competent raters) and three proficient (20% of all proficient raters) changed the sequence.

8.4 Length of Verbal Protocols

Written performance assessment research using verbal protocol analysis does not centre on differences in protocol length. Raters' individual differences in verbosity are mentioned by Wolfe et al. (1998) who conducted research into cognitive differences in rater behaviour. They observed differences in length of verbalisations and decided to compensate for individual differences in verbosity by converting protocol text length into ratio before looking into occurrences of behaviour types. The present study focuses on data gathered bearing in mind limitations of generating verbal protocols of heeded information (Ericsson & Simon, 1993). Therefore, any conclusion on language quality can only be tentative and needs further investigation.

Considering the research focus of the present study, compensation for verbosity does not seem to be necessary. Still, some differences regarding length can be observed. The verbal protocols in this study represent a wide variety in length. The 370 verbal protocols were different regarding word numbers: the shortest was produced by RR4 (635 words) and the longest by R8 (5,176 words) with a difference of 4,541 words, both raters belong to the proficient rater group (see Table 8.3 for details).

Table 8.3
Length of the Verbal Protocols in Word Numbers

	minimum	**maximum**	**mean**	**(sd)**
Competent raters	801 (R17)	4,542 (RR5)	2,299.82	99.70
Proficient raters	635 (RR4)	5,176 (R8)	2,772.40	294.16
Total	635 (RR4)	5,176 (R8)	2,791.41	671.75

Word counts of the verbal protocols show that competent raters produced shorter protocols (mean 2,299.82 words) than proficient raters (mean 2,772.40 words), but their protocols are closer to each other according to length, as the standard deviation is 99.70; whereas proficient raters' protocols were varied in length with a standard deviation of 294.16. The observation based on protocol length is that competent raters' protocols show less diversity in length than those of proficient raters.

Raters' rating performance may be influenced not only by the order they evaluate and award scores to the scripts, but by the time factor as well. Fatigue in attention may appear with time during rating that can influence rating performance. Thinking aloud when performing any task, such as rating, takes longer than without verbalising thoughts during task completion (Ericsson & Simon, 1993, pp. 249-251; Lumley, 2000). In addition, as findings of an earlier study by Congdon and McQueen show, differences in rater severity were detected during a large-scale rating exercise, which lasted nine days, and the differences in rating decisions affected the awarded scores. Raters had to re-rate scripts and the comparison of results provided evidence for raters' variance when rating a high number of scripts for a long time without being monitored (Congdon & McQueen, 2000).

As timing during rating was not recorded, comparisons can be made on the basis of protocol length only. Comparing the protocol lengths of competent and proficient raters, four tendencies could be identified: a descending curve, an ascending curve, a curve with little deviations, and a curve with jumps.

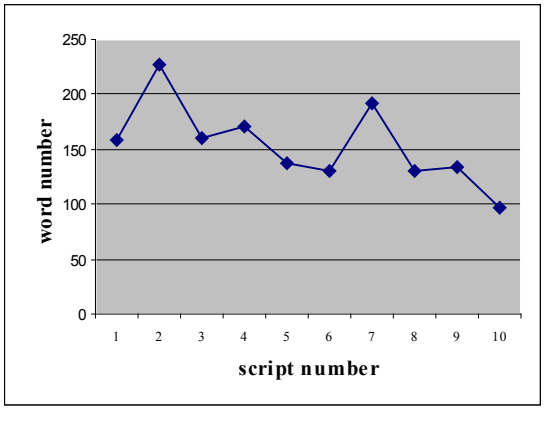


Figure 8.1. R11's protocol length (in number of words) during the rating sequence

The example of a competent rater, R11's protocol curve shows (see Figure 8.1) that she used more words when rating scripts at the beginning than at the

end of the rating task. Another example illustrates a more balanced tendency of a proficient rater, as looking at R6's protocol length (see Figure 8.2), we can see that her word number distribution curve is more balanced.

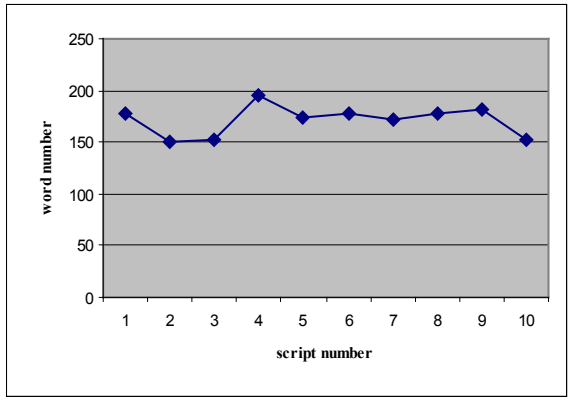


Figure 8.2. R6's protocol length (in number of words) during the rating sequence

There was only one rater of all 37 raters, whose protocol length curve was ascending; she was R18 from the competent group (see Figure 8.3).

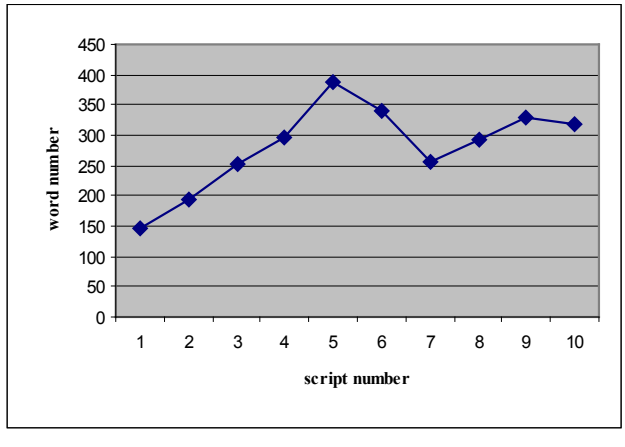


Figure 8.3. R18's protocol length (in number of words) during the rating sequence

Most raters' protocol length in both groups shows an uneven curve with higher and lower word numbers, as Figure 8.4 illustrates, in which a proficient rater's (RR6) protocol length changes over rating the ten scripts.

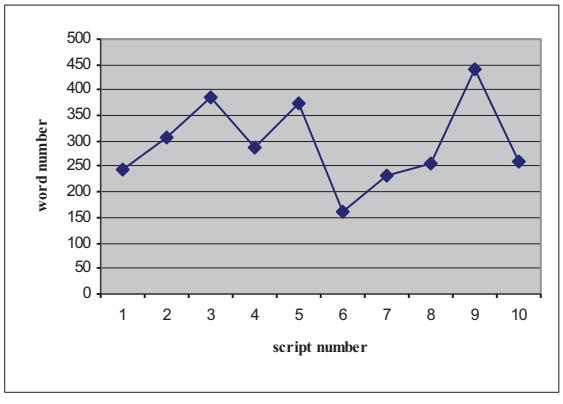


Figure 8.4. RR6's protocol length (in number of words) during the rating sequence

Eight (36%) competent raters' protocol length showed a descending curve, one uttered more words by the end of the rating process, 3 (14%) produced similar length texts, and 10 (45%) raters' protocol length curves show an uneven distribution in the number of words (see Table 8.4 for details). Proficient raters, on the other hand, produced different patterns as far as protocol length is concerned. Two out of 15 spoke less by the end than at the beginning, 4 (27%) produced protocols of similar in length, whereas 9 (60%) raters' protocols length curve did not show any tendency that can be related to sequencing.

Table 8.4
Protocol Word Number Distribution Curves

Curve type	Descending	Ascending	Even	Uneven
Competent raters	8 (36%)	1 (5%)	3 (14%)	10 (45%)
Proficient raters	2 (13%)	0	4 (27%)	9 (60%)
Total	10 (27%)	1 (3%)	7 (19%)	19 (51%)

Although data related to the time the raters spent rating is not available, examining the protocol length of competent and proficient raters, a tentative conclusion can be made: proficient raters' protocol length is more evenly distributed than that of competent raters, which may mean that their rating is less influenced by fatigue caused by time. The differences in protocol length can be attributed to other factors than the length of time spent rating them.

8.5 Raters' Rating Sequences

After examining some features of the raters' sequencing the scripts for rating and comparing length of protocols, rating sequences are investigated to get closer to patterns raters follow when rating. The observable occurrences of the raters' behaviour are first examined from the point of view of stages in rating sequences to see whether any patterns emerge from the verbal evidence of rater cognition. Thus, the research question inquiring into what characterises novice raters' rating processes means looking at sequencing rating stages to see whether there are identifiable patterns that raters follow. In order to find out more about patterns, the following more specific research questions can be posed:

- Do raters follow any kind of system across scripts and during rating each one?
- What jumps do raters make during rating?
- What identifiable patterns of behaviour do they follow in rating?

As to the next specific research question, which is closely related to the previous ones examining patterns that raters follow when rating written performance, the following question can be posed to examine the different patterns of rating behaviour:

- What patterns do raters follow when rating written performance?

The task seems to be ambitious, as humans can see the same phenomenon differently and even if they see it similarly, their thinking about it or the way they express their views can be totally different. This diversity is difficult to trace and research findings show that "readers appeared to lack a common language when talking about their judgements of essays" (Hamp-Lyons, 2007, p. 4). This finding does not seem to be encouraging and Hamp-Lyons goes on and says that raters "commented on the wide variety of approaches to judgements of essays" (2007, p. 4), and later, "readers seemed to have a strong but entirely personal set of expectations" (2007, p. 4). These findings imply that there are only tendencies that can be established when analysing rater behaviour, it does not seem possible to make generalisations.

As far as detecting any system in the raters' observable behaviour is concerned, two types of systematic behaviours are identified. One is a systematic way of following certain stages when rating over scripts, that is whether the raters employed a similar sequence of strategies for rating individual scripts. The other systematic behaviour is within rating the scripts, which means examination of the rating stages in each script, to see whether the raters followed similar sequences of strategies for individual scripts.

Frameworks of the rating processes discussed in detail in Chapter Four include several elements and their relationship, but they do not provide much evidence for the raters' actual rating sequence. The two most detailed models

by Milanovic et al. (1996) and Lumley (2000), however, suggest some sequence of rater behaviour. The model devised by Milanovic et al. (1996, p. 95) comprises seven stages: pre-marking, scanning, quick read, rating, modifying, revising and decision-making stages. Lumley (2000, p. 289) at the same time has a more complex model in which reading, scoring and conclusion stages are realised on three levels. At the beginning of rating, there is a so-called pre-scoring stage during which raters after the first reading comment on overall features of texts without assigning scores or specifying any of the rating criteria (Lumley, 2000). Analysis of the verbal data of the 37 raters of this study shows they demonstrate the behaviour types identified in the frameworks during rating and that there are some characteristics as far as systematic rating is concerned.

The leading rating principle for all raters seemed to be the four criteria as they appear in the rating scale, no matter how much attention was paid to the individual scale criteria, raters mainly followed the sequence of rating task achievement first, then vocabulary, then grammar, and organisation last. The example in Excerpt 8.1 illustrates systematic rating behaviour of a proficient rater's (RR2) rating sequence when she was dealing with script N1.

Excerpt 8.1: An example for systematic rater behaviour when rating a script

RR2 ***Script N1***

TU	Rater talk
1	**Thank your friend – yes Write about the journey home – yes Presents, write about presents – yes Invite your friend for next holiday – come here for next summer, yes Tell about programmes you are planning – go to the cinema, beach,** yes
2	Task achievement
3	we have one, two, three, four, five content points,
4	and it's an informal letter,
5	so that's 6
6	Vocabulary
7	let's see, *healing potion,*
8	*gipsy lady,*
9	quite a **wide range of words and expressions,**
10	it's a 5
11	And the next one is grammar
12	structures, **there are some inaccuracies,**
13	*like travelled for,*
14	*I felt as a fish at a* **(...),**
15	I don't know what that is,
16	*stand at the train,*
17	that's incorrect,
18	so **some inaccuracies, but the text is comprehensible**,
19	**there is a variety in structures,**
20	but there is more than one or two inaccuracies,
21	that's a 5,
22	grammar is 5
23	Now organisation
24	there is, thank you is one paragraph, journey is another paragraph, presents another paragraph, invitation, separate paragraph, and programmes, separate paragraph, there's a closing to the letter, there's a beginning,
25	**so it corresponds to the task,**
26	and **there's clear logical link,**
27	links are not very clear, especially here, at the end, as he talks about inviting the friend and then the programmes, doesn't (...),
28	organisation is a 5
29	6 plus 5 is 11, plus 10 is 21

RR2 first focused on task achievement from TU1 to TU5; then, she looked at the vocabulary announcing the criterion (TU6) which she rated until TU10.

Next, she identified the grammar criterion (TU11) and rated it and she awarded a score (TU22). Then, RR2 dealt with organisation and after identifying (TU23) it, she made five remarks (TU24 to TU28). Finally, she added up the scores (TU29).

The majority of the raters followed the order suggested by the rating scale systematically; however, there were some exceptions. There were 3 (14%) raters from the competent group who did not follow this sequence. One of them, R17, did not rate all aspects and RR6 evaluated more aspects at the same time, as the examples in Excerpt 8.2 show. R17 started with looking at task achievement, which she announced (TU1) and evaluated (TU2 to TU4) and then turned to organisation (TU5), identified the criterion (TU6), and chose a score (TU7). Next, she evaluated vocabulary: TU8, TU9 and TU10 refer to the vocabulary criterion. RR6 combined two rating criteria and announced the score saying that it was the same.

Excerpt 8.2: Examples for non-systematic rating sequences of R17 and RR6

R17 Script N3

TU	Rater talk
1	for task achievement
2	I gave the student 3 points
3	As all content points were mentioned
4	but not elaborated
5	I think, there are problems with coherence, link between the parts is not clear He did not connect them, so it is very drawling and not composition like
6	so for organisation
7	I gave 2 points
8	the same for vocabulary
9	2 points
10	because uses very few words compared to the level of the task and there is no variety

RR6 Script N4

TU	Rater talk
10	I gave 2 points for both grammar organization

There were two raters (13%) out of 15 in the proficient group who deviated from the sequence suggested in the rating scale, as the example from RR3's protocol in Excerpt 8.3 shows. As we can see, RR3 evaluated organisation (TU3 and TU4) before task achievement (TU5 and TU6) and after making an own focus comment (TU7) she went on and evaluated vocabulary (TU8). Then, she expressed her opinion on the writer's proficiency (TU9) and she rated grammar afterwards (TU10).

Excerpt 8.3: An example for non-systematic rating sequence of RR3

RR3 Script N7

TU	Rater talk
3	There is nice introduction, discussion, and everything is included
4	And there is logic between paragraphs
5	However the ending is missing
6	But solved with *I am waiting your answer*
7	It could have been made longer
8	Vocabulary is varied there are no repetitions
9	She is not absolutely sure about structures
10	10 Right in the second sentence we can see *Thank you to staying for weeks at your home*

The other feature that both competent and proficient raters shared was the diversity of rating sequences within each of the four criteria in the scale. However, there were some observable differences. There was no single rating sequence; each rater used several combinations of steps in rating each criterion. The difference was whether they employed the same two, three or four patterns during rating or they used a different pattern for each script and each rating criterion. They might first identify the rating criterion, chose a score, and then evaluated the script in their own words or read the matching score descriptor. Some raters applied different rating strategies for rating task achievement than for rating the rest and they stuck to such processes during rating. Some, however, used different sequences for each script. Competent raters showed less systematic behaviour than proficient ones, 7 raters (32%) out of 22 developed their own rating patterns and followed them throughout the whole rating process. The proficient group was more systematic in this respect: 10 raters (67%) out of 15 used similar sequences in rating.

To sum up, raters followed the rating sequence suggested by the rating scale order: first, they rated task achievement: then vocabulary, grammar, and organisation. This finding confirms what earlier studies reveal, "most of the time raters follow the rating categories provided, and in a very orderly way" (Lumley, 2002, p. 255). There were only some exceptions: few raters either randomly commented on the criteria or left some of them out or rated more at a time. As far as systematic rating behaviour is concerned, most competent and proficient raters followed rating criteria as presented in the rating scale systematically, however, there were raters in both groups who did not present any system in their rating processes. There were three raters (14%) in the competent group and two (13%) in the proficient one who did not demonstrate any systematic behaviour when dealing with individual scripts. Regarding rating steps within each script, raters showed more diversity: they used different sequences of

steps when dealing with the four rating criteria. Nevertheless, proficient raters presented more systematic sequences of stages than competent raters. The emerging patterns are examined in detail below after discussing pre-scoring stage and jumps in rating sequences.

The rating of each script started with script and criterion identification; however, there are many occurrences of overall comments at the beginning. Competent raters seemed to point out overall impression less frequently, 6 (27%) of them, whereas 7 raters (47%) from the proficient group started rating similarly. This stage in rating is identified in earlier research as pre-scoring stage during which raters make initial observations on text without special rating focus (Lumley, 2000; 2002; Milanovic et al., 1996). Raters either refer to surface features or to the content of the text. Surface features include handwriting, legibility or layout. One particular rater, R4 attempted to guess the gender of the writer according to handwriting, as an example in Excerpt 8.4 shows. She started rating each script similarly, guessing the writer's gender and she described the quality of script. Another rater, R14, commented the beginning regularly and said when rating script N9 that he found the script strange (TU2) and that there were missing parts (TU3), and then, explained that he meant legibility by "strange" in TU4.

Excerpt 8.4: An example of R4's comment on handwriting quality

R4 Script N9

TU	Rater talk
11	At first sight I always try to identify whether it is a woman's or man's handwriting and this time I was right, we can see man's handwriting, and what is quite frequent for boys in primary school is the way letters are formed

All these comments referred to the overall quality of the script before he stated rating the criteria one by one (see Excerpt 8.5). Similarly, when he rated another script, N4, made a point on legibility (TU2) which he developed into forming his overall impression on quality (TU3) and then he expressed his feeling (TU4).

Excerpt 8.5: Two examples of initial comments on overall quality

R14 Script N9

TU	Rater talk
1	Script number 9
2	It is rather strange for the first sight
3	The paragraphs are also missing
4	The writing is a bit strange, however you wouldn't consider it awful, it is eligible

R14 Script N4

TU	Rater talk
1	script number four
2	Huh, for the first sight very, very, very clumsy, and very bad handwriting
3	very strange the way it looks mmm... no letter format, I don't know my first impression is not too good
4	it seems rather important what you ... what your first impression is, yes, of course

The length of texts was assessed at the beginning, such as the example shown in Excerpt 8.6 in which R1 first established that the text was short (TU1) and then stated that it was difficult to read.

Excerpt 8.6: An example of initial comments on length and layout

R1 Script N2

TU	Rater talk
1	This composition is short, I think.
2	And it is difficult to read.

At the beginning raters sometimes referred to the content of scripts: they often expressed their feeling towards the text; one of them, R13 (see Excerpt 8.7) started rating by saying that she found the text funny (TU2), or RR4 expressed her opinion in connection with relevance (TU1). There were also remarks on overall quality at the beginning, such as TU1 in RR14' protocol illustrates.

Excerpt 8.7: Three examples of comments on text content

R13 Script N9

TU	Rater talk
1	Ok, script number 9
2	At least it was funny

RR4 Script N9

TU	Rater talk
1	The student didn't understand the task, and he didn't know what he was supposed to write about

RR4 Script N9

TU	Rater talk
1	This is a basically good letter also

To sum up rating sequences, raters did not refer to such strategies regularly, and even individual raters were not consistent in this respect. They did not seem to assign particular role to surface features and they very often started rating with the first criterion, which was task achievement. However, more proficient raters made overall comments when they started rating than competent ones. Now, before looking into observable rating patterns, let us see whether any jumps were observable in the rating processes.

The role the rating scale plays in rating scripts is apparent and most raters followed the order of the four criteria as they appeared. Raters attended to the four rating criteria one by one and went through the rating process either in a systematic or in a non-systematic way. The rating sequence was linear in this sense, however sometimes it was broken by jumps. As Wolfe observed, "less proficient scorers seem to *jump* from one content focus category to another as they evaluate an essay" (emphasis in original, 1997, p. 98). A jump means deviation from the criterion the rater focuses on. For example, while rating task achievement a rater talked about a grammar criterion and then, went back to rating task achievement. These jumps in rating processes indicated the raters' switch to a different criterion and their change of rating focus. While they attended to a criterion, they noticed something that related to another one and commented on it, as the example in Excerpt 8.8 shows.

Excerpt 8.8: An example of jumps in the rating process

R4 Script N4

TU	Rater talk
6	in spite of these achieves communicative goal
7	so the boy achieves what he wanted to say, although
8	I could give 5 points
9	Because it is somehow difficult, so it caused difficulty to read the text, the whole was awkward, clumsy, covered everything, all topics, so there were no problems with this criterion, the only thing is that it was illegible practically
10	There are many mistakes
11	There are scribbles, which influence layout negatively
12	Grammar mistakes, but let's go one by one

R4 when rating task achievement of script N4 identified the criterion first (TU6), expressed her opinion on task achievement (TU7) and chose the score (TU8). She provided an explanation for the decision (TU9) on task achievement which he supplemented with evaluating grammar (TU10) and layout (TU11) when she realised that she deviated from the rating sequence (TU12).

In this respect, examining the jumps in the rating processes, differences could be detected between competent and proficient raters' behaviour. Competent raters changed focus more often during rating a script than proficient ones. There were 10 (45%) competent raters who switched to a different criterion when rating and (27%) raters in the proficient group deviated from the criterion they attended to. These findings suggest that it is more typical of proficient raters to focus their attention on the criterion they are rating more than of competent raters. Next, I would like to look at possible patterns that raters employ in rating.

8.6 Raters' Rating Patterns

As the description of the diversity of sequencing rating processes shows, the raters did not follow uniform processes, they developed their own rating schedule which they employed in a systematic or a non-systematic way and they sometimes changed their focus while rating the individual criteria. It follows that rating processes are not only complex, but they also have different patterns. These patterns are difficult to outline, as they demonstrate the raters' individual characteristics and even individual raters do not follow the same pattern all the time. The two groups went through the rating processes in a similar way; the only observable differences between them were in the amount of the initial comments and the number of jumps they made. The proficient raters made more initial comments than competent ones, whereas the competent raters produced

more jumps in their rating processes. The patterns competent and proficient raters developed for rating were similar. Their different patterns are analysed below to reveal more about what the raters attended to with special focus on establishing the patterns of rating processes.

The rating of task achievement has some different features from the other three rating criteria: raters used different patterns for rating task achievement. They either considered the text as a whole and evaluated the content points, or segmented the text and read out the scale descriptor, or read the rubrics, or summarised the content. There is an example for treating the text as a whole and rating task achievement by RR11 (Excerpt 8.9). RR11 chose a score (TU1) after she identified the rating criterion of task achievement (TU2), commented on length (TU3). Then, she evaluated content points in her own words (TU4) and read out a scale descriptor (TU5). She mentioned style (TU6) and read out another scale descriptor (TU7).

Excerpt 8.9: An example of treating the text as a whole when rating task achievement

RR11 Script N3

TU	Rater talk
1	For the task achievement I gave
2	5 points,
3	because I think this letter is too short,
4	and there's limited information in it. The writer doesn't elaborate the points he's writing about.
5	Although **he achieves communicative goal** (it is by all means **a letter to a friend**),
6	the style is appropriate
7	and **he covers all 5 content points**.

The other pattern for rating task achievement was when raters broke the text up and evaluated content points included in the writing task one by one. There is an example in Excerpt 8.10 to illustrate how RR2 acted: first she read out a content point from the task rubric (TU1) and checked if it was discussed (TU2), next she went on quoting from the rubric (TU3) and read out the relevant part (TU4). TU5 was also a content point cited from the rubric followed by reading part of the text (TU6), and an evaluation of that content point (TU7). Next the content point from the rubric (TU8) was evaluated with own words (TU9) and, finally, task achievement was globally evaluated with a descriptor from the scale (TU10) and a comment on comprehension (TU11) was also made.

Excerpt 8.10: An example of breaking the text up for rating task achievement

RR2 Script N2

TU	Rater talk
1	OK, **thank your friend**
2	he hasn't got that
3	**Write about the journey home** – OK **Presents**
4	*pink pyjamas, Scottish whiskey,* yes
5	**Invite your friend for next holiday**
6	*furthermore one of my friends will make a party the next Friday, can you come,*
7	that's not an invitation for a holiday
8	**Tell about programmes you are planning**
9	there's none of that either
10	**It's a letter for a friend,**
11	I don't understand this bit here

A score was sometimes nominated before evaluation, as the example in Excerpt 8.9 shows, after criterion identification, or after evaluation. Score nomination patterns were different not only in rating task achievement, but for the other three rating criteria as well. Raters either nominated the score first and then evaluated the criterion or evaluated the criterion first and proposed a score afterwards. Both patterns are observable and individual raters had the same pattern throughout all scripts.

Rating patterns for vocabulary, grammar and organisation were similar, but the order of different features raters attended to was diverse. Raters started with criterion identification which was followed by reference to the scale either by simply reading a descriptor out or evaluating the criterion with own words. Some raters, as mentioned above, nominated a score first and then, as if they were justifying their choice, evaluation of the criterion followed. When evaluating a criterion, different rating behaviours occurred and raters read an example or examples from the script. Thus, the main identifiable patterns were as follows:

Pattern A
criterion identification → own focus → (example) → score

Pattern B
criterion identification → scale → (example) → own focus → score

Pattern C
criterion identification → score → (example) → own focus.

These patterns were often extended and the elements were repeated:

Pattern D
criterion identification → scale → example → scale → own focus → score

Pattern E
criterion identification → score → own focus → scale → own focus → (finalises) score

To sum up, rating patterns raters employ when evaluating individual criteria in the scale were diverse. Evaluation of the aspect of task achievement showed two different patterns: in one the text was considered as a whole, in the other the raters segmented the text and evaluated content points one by one. The rating patterns for the individual criteria were not linear, but iterative; raters went back to a strategy they employed before. Moreover, individual differences in patterns are difficult to identify; what seems observable is that there is no uniform pattern, even each rater developed more than one for attending individual rating criteria.

8.7 Conclusion

Rating processes are characterised by different sequencing features, some of which are discussed in this chapter. Sequencing can be related to the order in which each script is dealt with, to the order in which raters attend to different rating criteria, and also to the patterns of behaviour they develop during rating. First, features of sequencing scripts for rating were analysed to see how the order of scripts influenced rating behaviour. Although fatigue of attention is markedly observable in large scale testing context (Congdon & McQueen, 2000), a drop in word number by the end of the rating process in some of the protocols indicates that raters' attention may decrease with time. This tendency was more apparent for competent raters than for proficient ones. This observation is only a tentative one, as the number of scripts is low: still, the tendency that raters' attention drops by the end of rating process is informative and reported in earlier studies as well (Lumley, 2000).

As far as differences in protocol length are concerned, competent raters' protocols are more evenly distributed for each of the ten scripts than those of proficient raters. Extremely lengthy protocols were produced by the proficient raters; this finding can suggest that raters' proficiency is not influenced by the amount of language they use for evaluation. Overall length differences in verbal protocols were observed in previous studies and compensated for: Wolfe et al. (1998) attribute differences in protocol length to raters' differences in verbosity. The observation here seems to indicate that raters' individual differences in verbal protocol length do not influence rating performance. No observable relationship has been found between length and efficiency in rating. Most probably individual differences played an important role in how much text raters produced in the think-aloud protocols.

Raters go through the rating process governed by rating criteria either systematically following a linear process or non-systematically making jumps or leaving rating criteria out from their rating processes. Although regarding systematic behaviour no major differences were found, proficient raters seem

to be more organised in this respect, and their rating processes contained fewer jumps than competent raters' processes. Findings regarding attention to rating criteria are in line with earlier studies (Lumley, 2000; 2004; Wolfe, 1997). Some raters make overall comments initially without a special rating focus to express their initial observation related to text. This pre-scoring stage, however, does not appear in all raters' protocols, only 13 (35%) of all raters made comments without a scoring focus and more proficient raters than competent ones presented this behaviour type.

Rating patterns show individual features. Apart from individual differences, raters rating processes when rating any of the four rating criteria were mostly iterative; they very often altered rating behaviours and went back to one or two strategies several times. This finding was confirmed in an earlier research in which raters' rating patterns were found iterative when attending to rating criteria using an analytic rating scale (Lumley, 2000). Competent and proficient raters in this respect did not seem to have distinct characteristics, which suggests that rating performance does not depend on the sequences of behaviour that the 37 raters in the present study demonstrate. Differences between raters' behaviours and their rating processes need further examination from a different aspect. In what follows, raters' observable behaviours are examined in terms of what they pay attention to when rating and to what extent.

Chapter 9

Raters' Rating Focus

Introduction

In the previous chapter I focused on analysing different features of rating sequences and attempted to answer the first research question regarding features of the rating processes. This chapter investigates rater behaviour in further detail to answer the second research question by looking into what raters attend to while making their decisions:

2. What criteria do competent and proficient raters focus on in rating?

In order to be able to answer the question, first, raters' behaviour types are examined one by one and the differences between competent and proficient raters are established. Thus, the following specific research questions are posed:

What management, rating and reading strategies do raters use?
What other comments do they make?

It is assumed that the data collected on raters' cognition in their verbal protocols represent verbalisation of their thinking. There is no evidence that they can verbalise their thinking processes entirely (Ericsson & Simon, 1993); however, verbal protocols provide valuable data for observing different behaviour types, including raters' behaviour. To analyse raters' behaviour the notions "focus" and "observable behaviour" are used as they are employed elsewhere, for example by Lumley (2000, pp. 134-136), in his research where he attempted to reveal what raters attend to. "Focus" is what raters attend to during rating and "observable behaviour" is the evidence of the raters' thinking based on what they verbalise during rating. This taxonomy seems to be convenient in this study as well to trace what raters focused on while making their rating decisions and what they attended to.

Bearing in mind the coding scheme which aided data processing in Chapter Seven and the discussion of rating sequences in Chapter Eight, the present chapter follows the order of raters' comment types. They are sequenced as above without any intention of prioritising one type of rating behaviour over another. First, I analyse and discuss management strategies; then, rating categories to shed light on their focus during rating. Next, I look at reading strategies to

complete the analysis of how raters interpreted the script, the scale and the writing task rubrics. Finally, I explore comments reflecting the raters' own focus.

9.1 Raters' Rating Foci

In order to understand what raters focused on during rating, first, the categories of raters' comments are established. Data of the 370 verbal protocols were processed using a coding scheme described in Chapter Seven. Rater behaviour types are grouped into the following four broad categories: (1) management, (2) rating, (3) reading behaviours and (4) own focus comments.

Management behaviour comments refer to different aspects of the rating procedure, including script and score identification and remarks on the rating processes. They are collected in two separate groups for practical reasons, as explained in Chapter Seven: script and category identification in one, and all other strategies related to management of the rating process in another.

Rating behaviour category includes comments on any of the four rating aspects as they appear in the rating scale: task achievement, vocabulary, grammar, and organisation. As the coding scheme in Chapter Seven shows, they are combined forming eleven subcategories for each rating aspect.

Reading strategies are collected similarly, according to the four rating aspects and are subdivided into five subcategories each to see what the rater was reading: the rating scale, the script, an example, part of the text or rater summarised script content.

Own focus comments are those which raters made when they did not refer to any of the four rating aspects, but expressed their overall impression of the script, their feelings, speculated about the script quality, writers' intention and proficiency.

Looking at the total number of comments, we can see that there was a total of 10,980 (mean 297) on the ten scripts: fewer competent raters made a point 6,441 (mean 293) than proficient ones, who made a total of 4,539 (mean 303) remarks (see Table 9.1 for details). Raters focused on rating strategies most, there were 4,229 (mean 114) comments, their reading focus was the next with a total of 3,096 (mean 84) comments. Occurrences of management and own focus comments were much fewer: there were 1,773 (mean 48) and 1,882 (mean 51), respectively.

Table 9.1
Raters' Comments on the Four Rating Foci

	Management (mean)	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	1,027 (47)	2,443 (111)	2,053 (93)	918 (42)	6,441 (293)
Proficient	746 (50)	1,786 (119)	1,043 (70)	964 (64)	4,539 (303)
Total	1,773 (48)	4,229 (114)	3,096 (84)	1,882 (51)	10,980 (297)

However, the four rating foci show differences in the groups of competent and proficient raters. Figure 9.1 shows the mean distribution of comment occurrences on the four rating foci: competent and proficient raters' management focus was similar: 1,027 (mean 47) and 746 (mean 50) respectively. Rating focus was somewhat different, while competent raters made 2,443 (mean 111) comments, proficient raters made 1,786 (mean 119).

Figure 9.1. Distribution of mean comments of raters' foci

As far as reading focus is concerned, competent raters read more: 2,053 (mean 93) times than proficient raters 1,043 (mean 70). Regarding own focus, we can see that competent raters focused less on their criteria than proficient raters: competent raters referred 918 (mean 42) times and proficient raters 964 (64) times to own remarks.

Competent and proficient raters' foci show differences on the four categories of management, rating, reading and own focus and what follows is a detailed investigation into each of the categories to reveal more about the features of raters' rating processes.

9.1.1 Management Focus: Management Strategies

Comments with a management focus were grouped into two subgroups for practical reasons. Raters were asked to announce which script they were assessing and identify the rating aspect in the scale. Script and criterion identification have only informative value for establishing which script raters were dealing with and which criterion they were rating. In addition, they remarked on the rating procedure as well, and they repeated the score or added up the scores. When raters reasoning was not clearly identifiable I included it in this category together with what raters said on the rating process itself. Table 9.2 shows that competent raters announced script or criterion identification similarly to proficient ones, there were 785 (mean 36) and 537 (mean 36) comments, respectively. Proficient raters' observations were distributed less evenly (sd 4.90 and 5.82, respectively).

Table 9.2
Raters' Management Focus Comments

Code	**Behaviour type as appears in the coding scheme**	**Rater**	**Script number**											
			1	2	3	4	5	6	7	8	9	10	Total (mean)	(sd)
1	Identifies script/ criterion	Competent	87	80	82	76	74	78	69	82	79	78	785 (36)	4.90
		Proficient	63	55	56	48	48	60	56	48	57	46	537 (36)	5.82
		Total											1,322 (36)	
D	Other remarks on he rating processes	Competent	25	26	29	26	19	27	23	21	24	22	242 (11)	9.59
		Proficient	28	23	22	16	18	23	25	16	19	19	202 (13)	13.22
		Total											444 (12)	

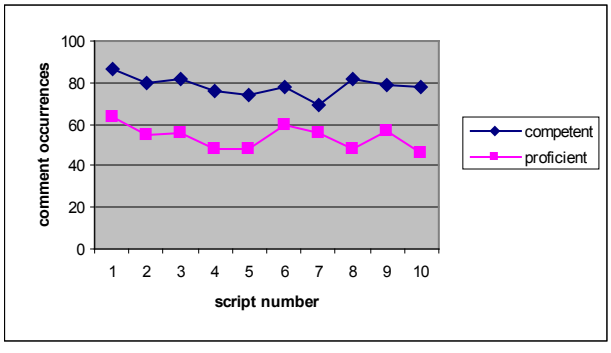


Figure 9.2. Comments on identifying a script or a criterion

There were differences in the number of comments for script and criterion identification for each of the ten scripts, as Figure 9.2 shows. This can be attributed to the fact that when raters changed their mind or repeated a score, or scores, they announced the aspect again, as in the example in Excerpt 9.1. When R7 completed rating a script, she repeated the scores, as we can see in the excerpt: in TU27 she announced the rating criterion and chose a score (TU28); then, identified the following rating criterion (TU29), said which score in TU30. Next, she changed focus (TU31) and proposed a different score (TU32) and explained which rating category it belonged to (TU33). Then, she repeated the score (TU34) and the rating criterion (TU35). For this script the occurrences of management behaviour comments were higher, as the rater repeated most of them.

The number of comments in the other group with management focus indicates that raters rarely expressed their opinion on the rating task they were accomplishing. Four subcategories were established for these strategies: when raters repeated the scores they assigned (code Da), when they added them up (code Db), when their reasoning was not clear (Dc), and when they evaluated the rating process itself (code Dd). Regarding comments on the rating process, competent raters made fewer remarks related to the rating process than proficient raters 242 (mean 11) and 202 (mean 13), respectively (see Table 9.2). Standard deviation for proficient raters' comments is significantly higher than for competent raters' (9.59 and 13.22, respectively).

Excerpt 9.1: R7's management comments

R7 Script N1

TU	Rater talk
27	for task achievement
28	I gave 4 points,
29	for vocabulary
30	3 points, no, sorry
31	sorry, sorry
32	5 points
33	for task achievement
34	4 points for
35	vocabulary

There are examples for each comment type in Excerpt 9.2. The four examples are from a competent rater's, R4's protocols rating four different scripts. The first one was score repetition at the end (TU47) of rating a criterion; the second was made at the end of rating script N9 when she added up the scores (TU44). The next example illustrates an unclear statement (TU8) and the fourth example shows how a competent rater, R4 noted what criterion she turned to.

Excerpt 9.2: R4's different management focus comments

R4 Script N5

TU	Rater talk	Code
47	for that I would give four points	Da

R4 Script N9

TU	Rater talk	Code
44	so altogether gets thirteen points	Db

R4 Script N3

TU	Rater talk	Code
8	here at the end, so ..... these enumerations	Dc

R4 Script N10

TU	Rater talk	Code
2	Now, then, I looked at the grammar mistakes the girl made	Dd

The distribution of comments on the rating processes for each of the ten scripts was different in some cases, as Figure 9.3 shows. The biggest difference was found between competent and proficient raters' comments for script N4: competent raters voiced their opinion 26 (mean 1.18) times, whereas proficient raters 16 (mean 1.07) times.

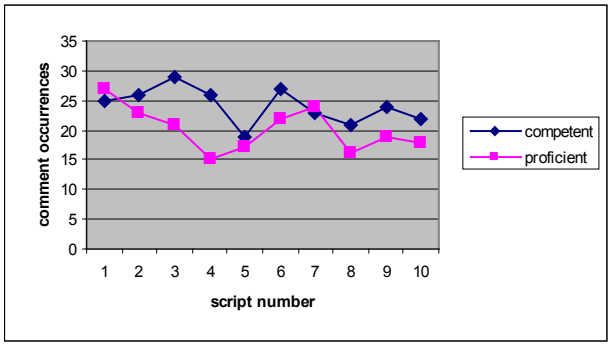


Figure 9.3. Comments on the rating process

Comment occurrences are summarised in Table 9.3 below and illustrate that both competent and proficient raters repeated the score and added them up on few occasions, and there were only two utterances with unclear reasoning.

Table 9.3
Management Focus Comments on the Rating Process for Script N4

Comment code	Da	Db	Dc	Dd	Total
Competent	3	12	2	9	26
Proficient	3	8	0	5	16
Total	6	20	2	14	42

Raters often reported and explained their proceeding in rating, as the examples from R13's, protocol show (see Excerpt 9. 3).

Excerpt 9.3: R13's management focus comments

R13 Script N4

TU	Rater talk	Code
6	Good	Dd

TU	TU Rater talk	Code
21	Good,	Dd

TU	TU Rater talk	Code
27	Two,	Da

TU	TU Rater talk	Code
33	So there at the second [script] it is nine altogether,	Db
34	Ok, for the third [script] then it is six and six adds up to twelve	Db
35	Ok, good... twelve	Da
36	Here then for number four [script] it is six and five that is eleven	Db
37	Good, ok	Dd

As we can see, R13, who was a competent rater, made several remarks on rating processes, the examples are taken from her protocol when she was rating script N4: in TU6, TU21 and TU37 she indicated completion of a rating step, she repeated the score twice (TU27 and TU35) and added up scores not only for the script she was dealing with (TU36), but for two scripts she rated earlier (TU33 and TU34).

To sum up, there were 1,322 comments (mean 36) found on script and rating criteria identification. This number clearly indicates that raters duly announced what exactly they were focusing on, which means that they attempted to follow the rating sequence as closely as possible. There were much fewer other management focus remarks, raters made a total of 444 comments (mean 12) in that category. Some raters repeated the score they chose and they sometimes added up the scores, although they were not instructed to do so. There were only few references to the rating process itself which were mainly technical statements, such as an indication of the following step in the rating process. An example in Excerpt 9.4 illustrates how a competent rater, R9 started the rating task. He first announced what he was doing (TU1), and then, identified the script he was dealing with (TU2) and indicated first reading (TU3).

Excerpt 9.4: R9's comments on rating

R9 Script 1

TU	Rater talk
1	Good afternoon. Welcome to "my introspective method", that is the thinking aloud procedure of the testing that I had to take.
2	Let's start with script number 1.
3	And after the first reading,

A proficient rater, RR8 made eight comments on the rating process, she announced six times that she was reading or re-reading a script and twice she talked about rating. She thought that script N2 did not conform to the task requirements and did not know what to do when rating task achievement.

Excerpt 9.5: RR8's comments on rating

RR8 Script N2

TU	TU Rater talk
17	If I give a zero, then I don't know if I should score vocabulary, grammar and organisation, that is, if I have to evaluate them at all. So if one aspect is zero can the rest be rated?

She expressed her doubt about rating (TU17), as in Excerpt 9.5. Raters rarely voiced difficulties, these findings may indicate that they could internalise the rating task successfully, and they did not often refer to problems in rating.

9.1.2 Rating Focus: Rating Strategies

Rating strategies were identified by calculating the occurrences of strategies during rating using the coding scheme. The strategies for the four categories, as they appear in the coding scheme are presented in Table 9.4 for each of the ten scripts and for both competent and proficient raters. Comparing the total means of comments raters made on the four aspects in the rating scale, it is apparent that they paid the most attention to grammar. Results show that both competent and proficient raters referred to grammar most frequently: 692 (mean 31) and 512 (mean 34) times, respectively.

Proficient raters paid the least attention to task achievement (374 comments; mean 25), whereas competent raters dealt with vocabulary the least (554 comments; mean 25). Standard deviation figures indicate the most uneven distribution of comments for rating task achievement by competent raters on the ten scripts (sd 12.54).

Table 9.4
Rating Focus Comments

Code	Behaviour type as appears in the coding scheme		Script number											
		Rater	1	2	3	4	5	6	7	8	9	10	Total (mean)	(sd)
A1	Task achievement	Competent	57	75	53	51	75	53	84	52	60	48	608 (28)	12.54
		Proficient	32	39	41	42	36	32	45	38	39	30	374 (25)	4.86
		Total											982 (27)	
A2	Vocabulary	Competent	59	54	49	55	56	60	47	54	66	54	554 (25)	5.42
		Proficient	55	54	40	40	48	42	46	45	54	40	464 (31)	6.11
		Total											1018 (28)	
A3	Grammar	Competent	60	65	76	82	72	65	66	61	78	67	692 (31)	7.44
		Proficient	55	48	61	54	48	45	55	49	51	46	512 (34)	4.98
		Total											1204 (33)	
A4	Organisation	Competent	52	61	62	67	62	53	57	57	64	54	589 (27)	5.04
		Proficient	47	46	40	36	41	38	49	45	47	47	436 (29)	4.48
		Total											1025 (28)	

Raters' foci on individual rating aspects were diverse: proficient raters focused less on rating task achievement and more on vocabulary, grammar and organisation. Rating vocabulary seemed to cause the most problems for the raters. When grouping raters into a competent and a proficient group according to their agreement with the benchmarks, as described in Chapter Ten, the scores awarded for vocabulary showed the least agreement among all raters. The correlation coefficients were the lowest on this rating aspect and there were only nine raters from the total group of 37, whose scores showed significant correlation with the benchmarks (see Table 7.4 for details). Comparing the findings from the two sources we can conclude that rating vocabulary was the most problematic; the agreement among raters was low although proficient raters paid more attention to this rating aspect than to task achievement and organisation.

As far as the distribution of rating strategies across the ten scripts is concerned, there were differences between competent and proficient raters for each rating aspect.

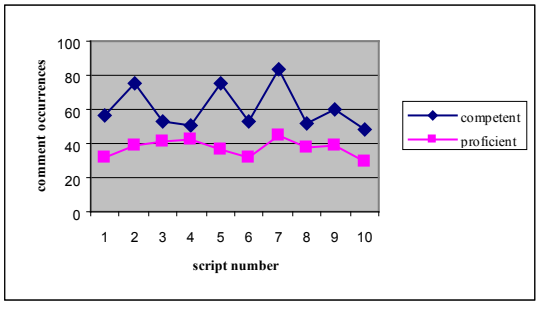


Figure 9.4. Comments when rating task achievement

Competent raters' evaluation of task achievement of individual scripts is uneven; they seem to pay unequal attention from script to script, as the curves in Figure 9.4 shows. Proficient raters, however, did not seem to be influenced by differences in script content, their attention to each of the ten scripts when rating task achievement was more balanced.

Comment occurrences on individual scripts when rating vocabulary demonstrated the most uneven spread among the four rating criteria. Although competent raters made fewer comments than proficient ones (mean 25 and 31, respectively), their comments were more evenly distributed across the ten scripts, whereas proficient raters' pattern was less consistent (see Figure 9.5 for details).

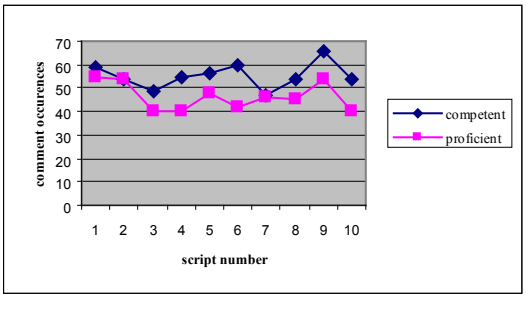


Figure 9.5. Comments when rating vocabulary

In addition, as far as proficient raters' comment distribution is concerned across the ten scripts, the highest variety (sd 6.11) occurred when dealing with vocabulary among the four rating criteria.

Raters' rating focus on grammar, as mentioned earlier, was the most intensive among the four rating criteria for both competent and proficient raters. The comment distribution curve for the ten scripts is less even for competent raters than for the proficient ones. As Figure 9.6 shows, the biggest differences in occurrences were in rating scripts N4 and N9: competent raters referred to grammar 82 (mean 3.7) and 78 (mean 3.5) times, respectively, whereas proficient raters considered grammar the most when they were rating script N3.

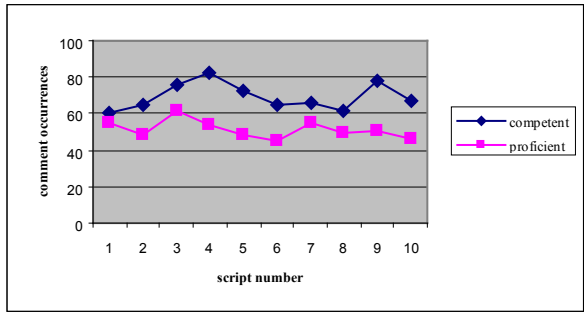


Figure 9.6. Comments when rating grammar

Finally, regarding organisation, which was the fourth rating aspect in the rating scale, raters made a total of 1,025 (mean 28) comments when rating the aspect. Competent raters focused less on organisation than proficient ones: they mentioned it 589 (mean 27) and 436 (mean 29) times, respectively and competent raters' comments were less evenly distributed (sd 5.04 and sd 4.48), as presented in Table 9.4. Distribution of rating focus remarks across the ten scripts when dealing with the aspect of organisation is illustrated in Figure 9.7.

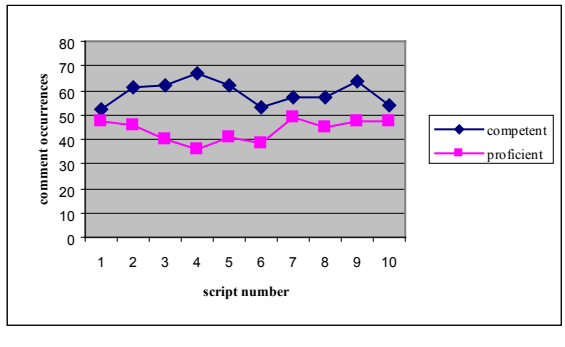


Figure 9.7. Comments when rating organisation

A comparison of rating comments made on the ten scripts by competent and proficient raters on organisation shows a big difference in occurrences for script N4: competent raters made 67 comments (mean 3), whereas proficient raters made 36 (mean 2.4) comments.

To sum up, raters paid considerable attention to each of the four rating criteria in the scale; however, there were some observable differences. Similarly to previous studies into rater behaviour (Lumley, 2000; Milanovic et al., 1995), raters seemed to prioritise linguistic features of written performance, and among them they were most concerned with structural accuracy. Findings of the study confirm this statement; raters paid more attention to grammar than to the three other rating aspects. Attention paid to rating task achievement for competent raters seemed to be uneven, the occurrences of rating comments on individual scripts was varied. Overall, competent raters made more comments when rating task achievement than when rating vocabulary and organisation. Proficient raters' comment occurrences were higher than competent raters' when rating vocabulary, grammar and organisation. Proficient raters showed more attention to the aspects of vocabulary, grammar and organisation than to rating task achievement.

9.1.3 Reading Focus: Reading Strategies

The five different reading strategies identified during rating the four aspects: task achievement, vocabulary, grammar and organisation were: reading the scale, the script, or the task rubric. In addition, if raters quoted an example or summarised the script content, the comment was also classified as reading behaviour.

Table 9.5
Reading Focus Comments

Code	Behaviour type as appears in the coding scheme	Rater	Script number											
			1	2	3	4	5	6	7	8	9	10	Total (mean)	(sd)
B1	Task achievement	Competent	76	83	72	79	82	59	62	66	68	58	705 (32)	9.31
		Proficient	22	15	21	15	20	21	20	15	23	23	195 (13)	3.27
		Total											900 (24)	
B2	Vocabulary	Competent	56	64	53	55	41	36	52	27	70	36	490 (22)	13.59
		Proficient	32	49	22	37	21	39	38	28	34	40	340 (23)	8.59
		Total											830 (22)	
B3	Grammar	Competent	49	40	65	81	80	49	56	70	57	42	589 (27)	14.70
		Proficient	43	30	40	56	31	31	42	29	32	31	365 (24)	8.66
		Total											954 (26)	
B4	Organisation	Competent	25	36	28	31	27	28	23	21	23	27	269 (12)	4.36
		Proficient	14	27	15	11	8	14	13	11	12	18	143 (10)	5.21
		Total											412 (11)	

Occurrences of reading focus comments show a varied picture on the four rating aspects. Similarly to rating focus, occurrences of reading focus comments demonstrate that raters paid the most attention to the aspect of grammar: as Table 9.5 shows, they referred to reading strategies 954 (mean 26) times when rating grammar. However, comparing competent and proficient raters' reading focus, competent raters turned to reading strategies more often when dealing with task achievement than when rating any other aspect. Their reading focus comments on task achievement: 705 (mean 32), outnumber any other occurrences. Reading strategy was the least used when rating organisation: competent raters remarked on organisation 269 (mean 12) times, while proficient raters fewer times (143, mean 10).

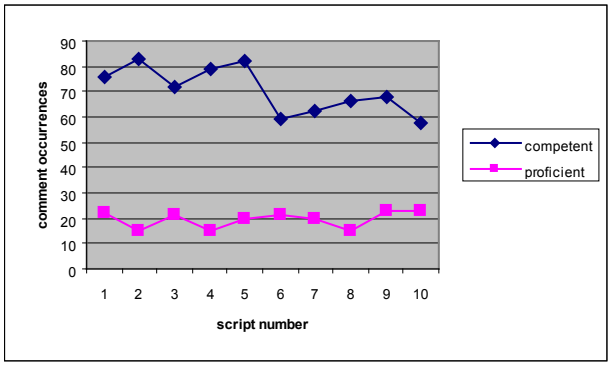


Figure 9.8. Comments when rating task achievement

As mentioned above, there were big differences in reading strategy use between competent and proficient raters. On the one hand, this difference could be observed in the occurrences of reading focus related comments. On the other hand, when looking at the comment distribution for the ten scripts, as demonstrated in Figure 9.8, comments were more evenly distributed for proficient raters. Competent raters' comment distribution curve is descending: they seemed to turn to reading strategies less frequently by the end of the rating process.

Although both competent and proficient raters' focus on vocabulary was similar (means 22 and 23, respectively), comment distribution across the ten scripts was very varied, as shown in Figure 9.9.

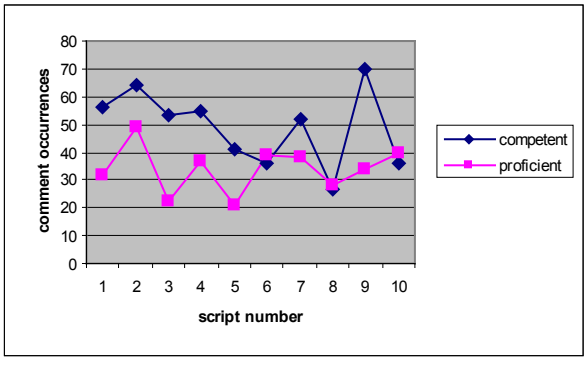


Figure 9.9. Comments when rating vocabulary

Reading strategies are spread unevenly for both groups: competent raters read more when rating script N9 than when rating any other script and they read the least when rating script N8. Proficient raters turned to reading the most when rating script N2 and the least when rating the aspect of vocabulary of scripts N3 and N5.

Looking at the distribution of comments related to reading behaviour in rating grammar we can see that competent raters' comment distribution is less even than that of proficient raters. As Figure 9.10 illustrates, competent raters employed the most reading strategies when dealing with script N4, N5 and N8.

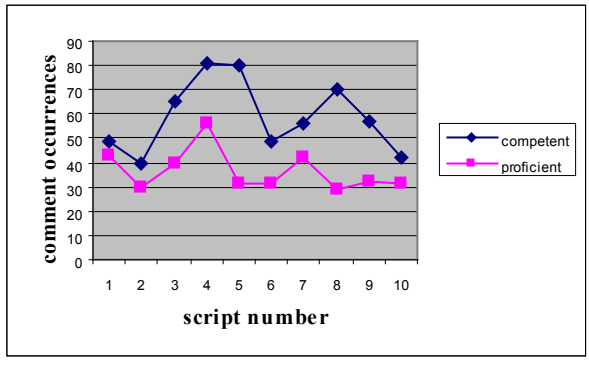


Figure 9.10. Reading focus when rating grammar

Proficient raters paid similar attention to script N4, whereas their comment occurrences on scripts N5 and N8 were among the fewest when rating vocabulary of the other scripts.

Regarding rating aspect of organisation, there were the fewest reading strategies used by both competent and proficient raters (means 12 and 10, respectively). Distribution of comments across the ten scripts shows similarities, as the curves in Figure 9.11 show, the two curves have similar tendencies.

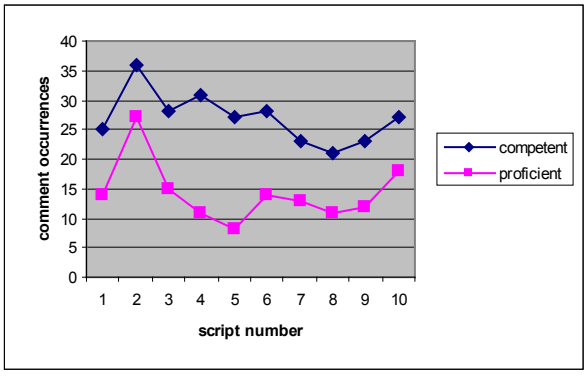


Figure 9.11. Reading related comments when rating organisation

To sum up observations on reading behaviour comments, we could see that by and large raters paid more attention to the aspect of grammar than to any other aspect, similarly to rating behaviour comments, but there were differences when looking at competent and proficient raters' reading focus. Competent raters made the most reading behaviour-related comments when dealing with task achievement, for them the aspect of grammar was the second most frequent. On the other hand, proficient raters paid much less attention to task achievement; they made more comments on grammar and vocabulary than on task achievement. Raters paid the least attention to the aspect of organisation, this aspect seems to be rated with few reading strategies.

9.1.4 Raters' Own Focus: Other Comments

Raters' rating processes mainly followed the four rating aspects of task achievement, vocabulary, grammar and organisation. There were instances of deviation from these criteria though, and raters focused on features of scripts that were not included in the rating scales or they meditated on writers' intention and proficiency. In addition, they sometimes corrected errors in texts or suggested task solutions. These strategies were collected in a separate group with the label "Other comments (own focus)" (see Figure 7.4 for details) in the coding scheme to make observations on raters' deviation from the assigned criteria. There were thirteen subcategories identified and as they represent different focus, these comments are discussed one by one as follows without an intention of prioritising any type over the other.

The first subcategory in the coding scheme is raters' reflection on script length: these comments most often appeared at the pre-scoring stage, as discussed in Chapter Eight. Raters' reflection on script length was not frequent, competent

and proficient raters made the same number of comments: 54 (mean 2.5) and 37 (mean 2.5) in this category, as shown in Table 9.6.

Table 9.6
Comments on Script Length

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Ca	Reflects on length	Competent		3	2	28	2	3	3	3	3	5	2	54 (2.5)	7.99
		Proficient		6	2	20	0	1	2	1	1	1	3	37 (2.5)	5.96
		Total		9	4	48	2	4	5	4	4	6	5	91 (2.5)	

The number of occurrences across the ten scripts shows that not all raters reflected on each script length, however, script N3 did not conform to raters' expectations as far as length is concerned, as the comment distribution curve in Figure 9.12 demonstrates.

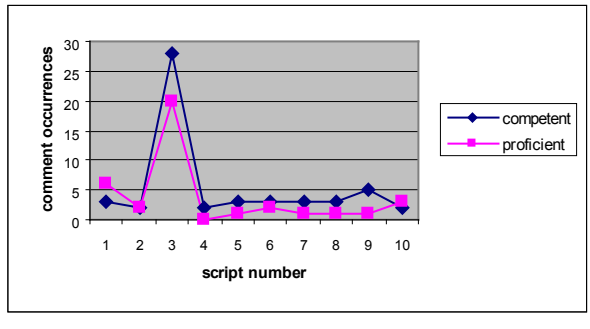


Figure 9.12. Comments on script length

Competent and proficient raters evaluated script length similarly; they all expressed their dissatisfaction with the length: competent raters made 28 (mean 1.3) and proficient raters 20 (mean 1.3) comments on length of script N3. The example of a competent and a proficient raters' evaluation of script N3 in Excerpt 9.6 illustrates how raters interpreted length deficiency for script N3.

Excerpt 9.6: R6's and R7's comments on script length

R6 Script N3

TU	Rater talk
3	I find this letter rather short, so it cannot be 150 words altogether

R7 Script N3

TU	Rater talk
3	Uhm, too little is written, the letter is too short at first sight

The competent and the proficient rater made a remark on length at the beginning of rating, TU2 and TU3, R7 made a general statement saying that the script was short, whereas the proficient one referred to the task requirement of at least 150 words.

The second subcategory comprised script eligibility and tidiness, which was not always considered. There were more comments made in this subcategory than for length characteristics, a total of 127 (mean 3.4) as it appears in Table 9.7.

Table 9.7
Comments on Script Eligibility and Tidiness

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cb	Reflects on eligibility	Competent		3	9	4	31	1	3	3	6	4	2	66 (3)	8.86
		Proficient		7	10	1	20	3	2	4	4	6	4	61 (4.1)	5.53
	tidiness	Total		10	19	5	51	4	5	7	10	10	6	127 (3.4)	

Proficient raters' mean was higher regarding remarks on script legibility and tidiness (61, mean 4.1) than competent raters' (66, mean 3), but proficient raters' comments were more evenly distributed: standard deviation is 5.53 for proficient raters' and 8.86 for competent raters' comments.

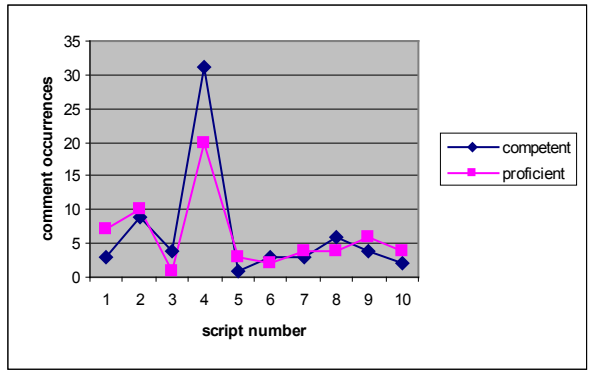


Figure 9.13. Comments on eligibility and tidiness

Raters did not remark frequently on script eligibility and tidiness; however, looking at the distribution curves in Figure 9.13, we can see that script N4 generated several comments from both competent and proficient raters. They commented on eligibility saying that the script was very difficult to read, as the two examples from a competent and a proficient rater illustrate in Excerpt 9.7.

Excerpt 9.7: R1's and R2's comments on script eligibility and tidiness

R1 Script N4

TU	Rater talk
1	This composition is very difficult to read, the writer has an illegible writing and it is messy

R2 Script N4

TU	Rater talk
44	organisation, there are many crossed out text parts, overwriting, arrows, after all it is not always legible

The competent rater, R1, first mentioned eligibility (TU1) and she added that the handwriting was very messy. The proficient rater's, (R2) observation on layout features appeared when she was dealing with the aspect of organisation in TU44.

Occurrence patterns for the comments on eligibility and tidiness of the other nine scripts were similar, competent and proficient raters seemed to pay similar amount of attention to these surface features.

Table 9.8
Comments on Script Quality

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cc	Reflects on script quality	Competent		5	2	2	2	2	5	2	2	1	4	27(1.2)	1.42
		Proficient		2	1	0	1	5	4	2	5	2	6	28 (1.9)	2.04
		Total		7	3	2	3	7	9	4	7	3	10	55 (1.5)	

The following comment type was script quality. Only few raters reported on quality: there were a total of 55 (mean 1.5) comments in this subcategory. However, proficient raters mentioned script quality more often than competent raters: 28 (mean 1.9) and 27 (mean 1.2), respectively (see Table 9.8 for details).

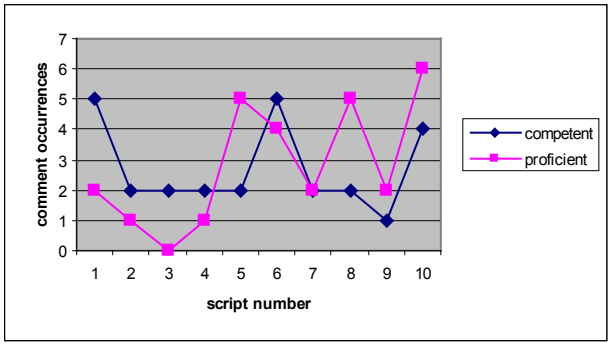


Figure 9.14. Comments on script quality

The two curves in Figure 9.14 are different regarding competent and proficient raters' comment distribution on script quality; the biggest differences were found on scripts N1 and N5. There are two examples of what raters said on script quality in Excerpt 9.8 to illustrate that competent and proficient raters' opinion was similar even if occurrences showed different patterns.

Excerpt 9.8: R7's and RR3's comments on script quality

R7 Script N1

TU	Rater talk
3	I think this sentence is good

RR3 Script N1

TU	Rater talk
18	But I still think, regardless of that that it is a pretty good composition, confirms requirements

When rating script N1, a competent rater, R7, mentioned quality at the beginning: in TU3 she said that the sentence she was reading was good; RR3, a proficient rater, made a similar comment later (TU18).

Table 9.9
Comments on Overall Impression

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cd	Comments on overall impression	Competent		4	9	5	7	3	13	4	1	4	3	53(2.4)	3.50
		Proficient		3	4	5	5	5	7	4	9	11	6	59(3.9)	2.47
		Total		7	13	10	12	8	20	8	10	15	9	112(3)	

There were more remarks on overall impression by proficient raters 59 (mean 3.9) than by competent raters 53 (mean 2.4) and proficient raters' comments were more evenly distributed across the ten scripts (sd 3.5 and sd 2.47 respectively) as demonstrated in Table 9.9.

Looking at competent and proficient raters' comment distribution on overall impression (Figure 9.15), there are big differences: comparing comments on script N6 and N9, we can see that competent raters made more remarks on the former and fewer on the latter, whereas the tendency is the opposite for proficient raters.

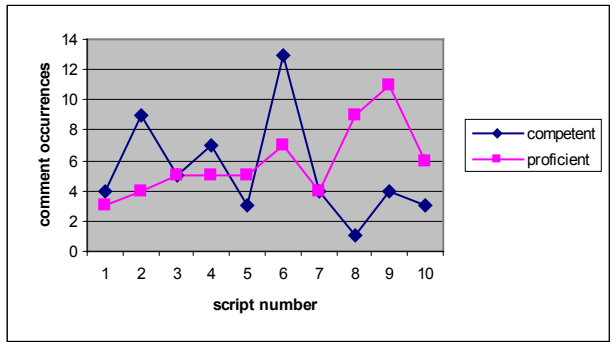


Figure 9.15. Comments on overall impression

The difference between a competent and a proficient remark regarding the first impression on script N6 (see Excerpt 9.9) is that the competent rater commented on overall impression at the end of rating the script (TU30), whereas the proficient rater at the beginning (TU2).

Excerpt 9.9: R7's and R14's overall impression comments on script N6

R7 Script N6

TU	Rater talk
30	But otherwise, my impression is not bad about it

R14 Script N6

TU	Rater talk
2	Hm... for the first reading it seems to be very goodt

The occurrences of comments for script N9 show the reverse tendency: proficient raters commented more on overall impression. Excerpt 9.10 is an example how raters expressed their overall impression: a competent rater, R4, said that the text caused some comprehension problems but she understood what the writer wanted to say. R9, a proficient rater said that although the task was completed (TU1), she found problems with content (TU2).

Excerpt 9.10: R4's and R9's overall impression comments on script N9

R4 Script N9

TU	Rater talk
2	So after reading the text, comprehension caused some problems here an there, but the message came through alright

R9 Script N9

TU	Rater talk
1	After all, he completed the task
2	However the whole letter looks a bit chaotic for some reason

Raters mentioned text comprehensibility also, competent raters remarked more, (93; mean 4.2) than proficient ones (51; mean 3.4), as Table 9.10 shows.

Table 9.10
Comments on Text Comprehensibility

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Ce	Comments on comprehensibility	Competent		16	19	3	14	6	8	2	5	15	5	93 (4.2)	6.11
		Proficient		11	6	3	7	1	4	3	3	12	1	51 (3.4)	3.87
		Total		27	25	6	21	7	12	5	8	27	6	144 (3.9)	

Competent and proficient raters talked about text comprehensibility across the ten scripts differently, as standard deviation figures show (sd 6.11 and sd 3.87 respectively). However, the comment distribution curves are similar in tendency except for script N2 (see Figure 9. 16).

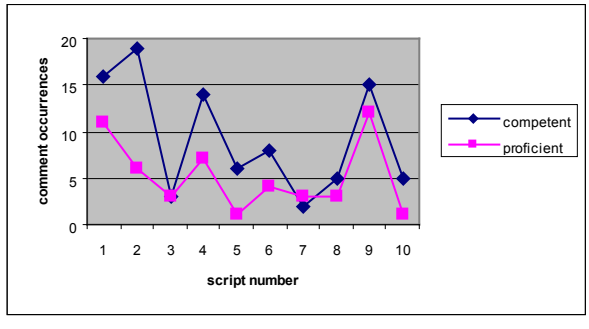


Figure 9.16. Comments on text comprehensibility

When rating script N2, competent raters made 19 comments (mean .9) and proficient raters 6 (mean .4), there are examples from a competent rater and a proficient in Excerpt 9.11.

Excerpt 9.11:R8's and R16's comments on text comprehensibility of script N2

R8 Script N2

TU	Rater talk
1	but at the end it is totally unclear for me what he is talking about, where he moved and where he is now

R16 Script N2

TU	Rater talk
5	I think, it's not really comprehensible

The two comments were made at the beginning of rating process, a proficient rater (R8) reflected on comprehensibility in TU3 and a competent rater (R16) reacted similarly in TU5, in addition R8 specified her comprehension problem and said that she could not make out what the writer intended to say.

Raters sometimes evaluated students' proficiency: proficient raters elaborated on their observation of student proficiency in more cases than competent raters. Competent raters made 34 comments (mean 1.5) and proficient raters 46 (3.1) in this respect (see Table 9.11).

Table 9.11
Comments on Observation of Student Proficiency

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cf	Comments on observation of student proficiency	Competent		1	8	1	6	2	6	2	1	3	4	34 (1.5)	3.87
		Proficient		4	6	3	1	4	3	7	8	5	5	46 (3.1)	2.07
		Total		5	14	4	7	6	9	9	9	8	9	80 (2.2)	

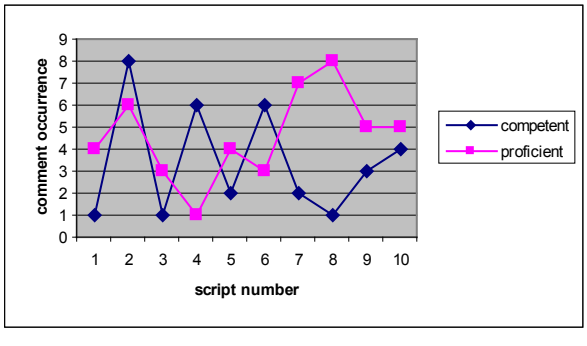


Figure 9.17. Comments of student proficiency

A comparison of the two curves of comment distribution in Figure 9.17 illustrates that competent and proficient raters reacted differently to student proficiency features and that different scripts generated a different number of comments in this respect.

Excerpt 9.12: RR3's and RR8's comments on student proficiency

RR3 Script N8

TU	Rater talk
12	uses grammar structures which show that he is confident in language use

RR8 Script N8

TU	Rater talk
13	But at least attempted, it means that knows it [the expression]

TU	Rater talk
17	According to these there are some uncertainties

One script, N8, made raters speculate on student's proficiency: proficient raters referred 8 times to the writer's language knowledge (see Excerpt 9.12). However, raters' evaluation was not similar: RR3 saw structure use as a sign of confidence in language use (TU12), while RR8 identified a language error and considered it as an attempt (TU13) and later she attributed mistakes to lack of confidence (TU17).

Although error correction was not included in the rating task, raters frequently corrected errors. Competent raters corrected errors in fewer cases than proficient raters: they made 157 (mean 7.1) and 123 (mean 8.2) comments, respectively, in this subcategory (see Table 9.12 for details.

Table 9.12
Error Correction Comments

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cg	Corrects an error	Competent		27	14	13	16	26	5	10	18	17	11	157 (7.1)	6.83
		Proficient		23	12	5	9	18	10	7	10	20	9	123 (8.2)	5.96
		Total		50	26	18	25	44	15	17	28	37	20	280 (7.6)	

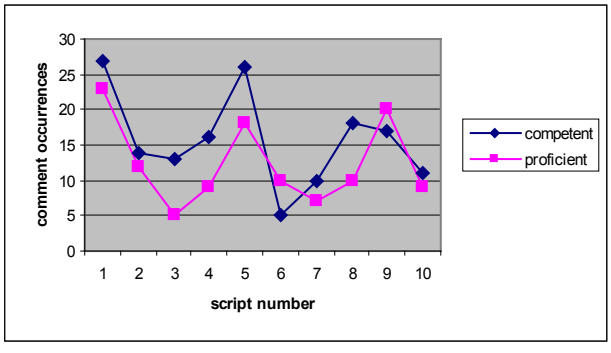


Figure 9.18. Error correction comment occurrences

The distribution of comments related to error correction is similar in tendency across the ten scripts (see Figure 9.18 for details); however, proficient raters reflected more on errors and competent raters' comments were less evenly distributed. The most errors were corrected when rating script N1: competent raters corrected 27 times (mean 1.2) and proficient raters 23 times (mean 1.5).

Table 9.13
Reflection on Own Feeling Comments

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Ch	Reflects on own feeling	Competent		20	26	19	29	23	25	13	17	16	23	211 (9.6)	4.98
		Proficient		24	42	20	32	31	44	28	35	27	29	312 (20.8)	7.50
		Total		44	68	39	61	54	69	41	52	43	52	523 (14.1)	

Raters very often expressed their feelings during rating: the occurrences of raters' own feeling comments were the highest among all; 523 (mean 14.1), as in Table 9.13).

Proficient raters reported their feelings considerably more frequently: 312 (mean 20.8; sd 7.50), whereas competent raters made 211 (mean 9.6 sd 4.98) comments expressing their own feeling.

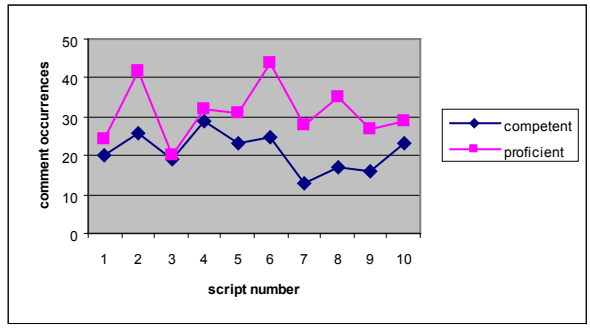


Figure 9.19. Comments on own feeling

As far as the distribution of comments across the ten scripts is concerned, competent and proficient raters' curves are similar in tendency except for script N2, on which competent raters made 26 (mean 1.2) and proficient raters made 42 (mean 2.8) comments (see Figure 9.19 for details). Remarks on content relevance were infrequent: 68, mean 1.8, as Table 9.14 shows.

Table 9.14
Comments on Content Relevance

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Ci	Reflects on relevance of content	Competent		1	11	2	3	6	1	4	0	8	2	38 (1.7)	3.52
		Proficient		5	6	2	1	5	0	5	1	5	0	30 (2.0)	2.40
		Total		6	17	4	4	11	1	9	1	13	2	68 (1.8)	

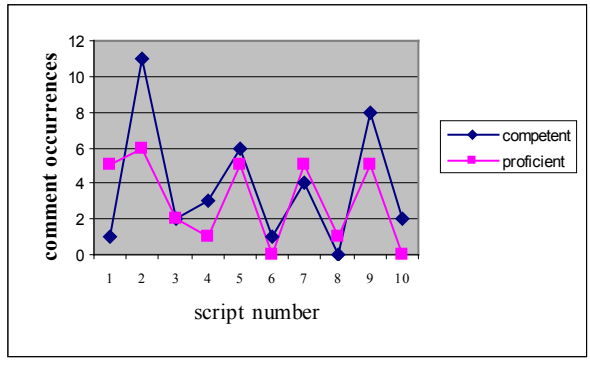


Figure 9.20. Comments on content relevance

The discussion of content relevance of each script was similar except for scripts N1 and N2, where there were big differences between comment occurrences: one competent rater expressed her concern of relevance of script N1, whereas there were five comments made on the same script by proficient raters (see Figure 9.20 for details). The comment ratio was the opposite for script N2, there were 11 competent raters and 6 proficient ones who referred to relevance.

Excerpt 9.13: R5's and R2's comments on text relevance

R5 Script N1

TU	Rater talk
9	The next one, the invitation is not a real one

R2 Script N1

TU	Rater talk
33	This somehow doesn't fit here, we don't even understand what the student is talking about

Looking at the two examples in Excerpt 9.13 when raters talked about text relevance, we can see that a competent rater, R5, mentioned one of the content points when rating task achievement (TU9), and a proficient rater, R2, referred to irrelevance in connection with vocabulary (TU33). Raters in some cases attempted to guess what the student's intention was; there were 86 (mean 2.3) occurrences of this comment type in the protocols. Competent raters made fewer such comments, 36 (1.6) than proficient raters, who made 50 (mean 3.3) remarks on student intention as Table 9.15 shows.

Table 9.15
Comments on Student Intention

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cj	Meditates on student intention	Competent		2	5	0	0	4	5	5	4	10	1	36 (1.6)	3.03
		Proficient		7	4	1	5	3	3	3	8	9	7	50 (3.3)	2.62
		Total		9	9	1	5	7	8	8	12	19	8	86 (2.3)	

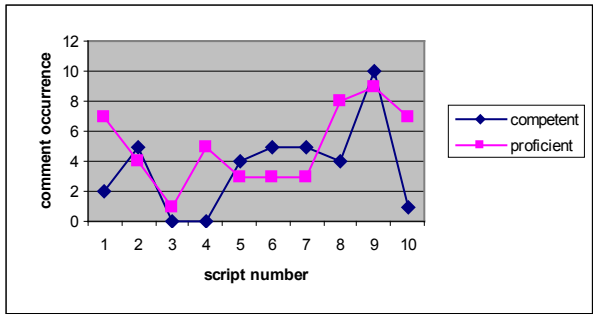


Figure 9.21. Comments on student intention

The distribution of occurrences of such comments shows differences between competent and proficient raters: the two curves are dissimilar (see Figure 9.21). However, the content of the remarks was similar, as illustrated in Excerpt 9.14, which illustrate raters' attempt to comprehend texts.

Excerpt 9.14: R4's and R2's comments on student intention

R4 Script N1

TU	Rater talk
10	and now I have to think what s/he wanted to say and it is because I can translate from Hungarian into English what s/he imagined

R2 Script N1

TU	Rater talk
3	It seems as if the writer refers to a lot of things, those s/he imagined, or maybe these happened on the trip

A competent rater, R4, was rating task achievement and meditated on student intention (TU10) and she tried to make sense of the text by translating ideas. A proficient rater, R2, interpreted the content as if it was imagination not reality (TU3) at the beginning of rating process.

Table 9.16
Suggestions on Solution

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Ck	Makes suggestion on solution	Competent		3	3	8	10	5	2	5	12	13	2	63 (2.9)	4.16
		Proficient		3	4	13	9	9	5	15	11	14	1	84 (5.6)	4.93
		Total		6	7	21	19	14	7	20	23	27	3	147 (4)	

Raters frequently provided a possible solution, as Table 9.16 shows, competent raters made much fewer such comments, 63 (mean 2.9) than proficient raters, who offered a solution on 84 (mean 5.6) occasions. The distribution of comment occurrences is uneven (see Figure 9.22), while competent raters gave a solution the most when rating scripts N4, N8 and N9, proficient raters did so for scripts N3, N7 and N9.

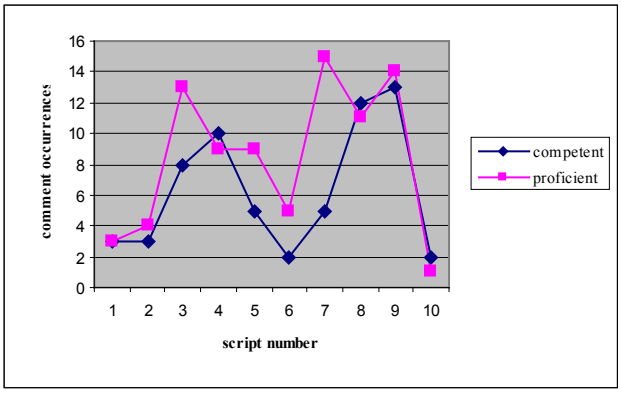


Figure 9.22. Raters' suggestions on solution

The suggestion a competent rater, R7, made, shown in the example in Excerpt 9.15, referred to word use (TU5) and RR13 proposed paragraph use in TU9 when they were rating script N9.

Excerpt 9.15: R7's and RR13's suggestions on solutions

R7 Script N9

TU	Rater talk
5	Something else should have been written here, "honest" doesn't make sense here, a totally different word should have been used, let's say "I am very happy" etc

RR13 Script N9

TU	Rater talk
9	S/he should have paid more attention to such minute details, and perhaps come up with paragraphs

Sometimes raters compared scripts they were rating to a previously rated script or more scripts and they occasionally compared scores to each other (see Table 9.17). Proficient raters turned to this strategy more often than competent raters: competent raters made 58 (mean 2.6) comments and proficient raters 52 (mean 3.5) comments respectively.

Table 9.17
Comments When Comparing to Other Script or Score

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
CI	Compares to other script/ score	Competent		3	2	4	4	7	11	8	6	7	6	58 (2.6)	2.66
		Proficient		1	0	5	7	6	3	7	11	3	9	52 (3.5)	3.49
		Total		4	2	9	11	13	14	15	17	10	15	110 (3)	

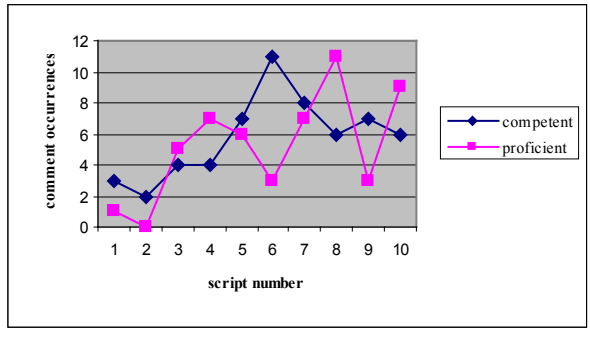


Figure 9.23. Comments when comparing to other script or score

Distribution of comments across the ten scripts is uneven for both competent and proficient raters and there are clear differences between the two groups as well, as Figure 9.23 shows. For example, when rating script N6, competent raters made eleven comments, whereas proficient raters made three.

Excerpt 9.16: RR13's and RR14's comparing scripts to each other

RR13 Script N6

TU	Rater talk
57	As it [clear logical link] didn't occur in the previous ones either

RR14 Script N6

TU	Rater talk
1	I found this one the best among all scripts I had to evaluate

The comment a competent rater (RR13) made on script N6 was a negative one (see Excerpt 9.16), referring to a lack of logic in the text (TU57) that she found missing from other scripts as well; a proficient rater RR14, in contrast, said that it was the best script she had read (TU1).

Finally, comments expressing uncertainty are investigated: raters said 59 (1.6) times that they hesitated when rating a script. Competent raters made fewer comments in this subcategory, 28 (mean 1.3) than proficient raters, who made 31 (mean 2.2) comments, as shown in Table 9.18.

Table 9.18
Expression of Uncertainty

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
Cm	Expresses uncertainty	Competent		1	4	4	1	4	4	3	5	1	1	28 (1.3)	1.62
		Proficient		3	6	2	6	1	3	2	3	4	1	31 (2.2)	1.79
		Total		4	10	6	7	5	7	5	8	5	2	59 (1.6)	

There are differences between competent and proficient raters' comment distributions on individual scripts, as Figure 9.24 illustrates, especially when rating script N4, one competent rater made a comment on uncertainty, while there were six comments made by proficient raters in the same subcategory.

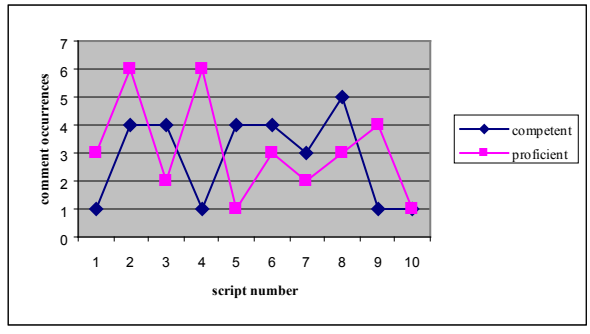


Figure 9.24. Comments expressing uncertainty

Raters, as the two examples in Excerpt 9.17 illustrate, expressed their uncertainty similarly: a competent rater, R4, said the same (TU19) as a proficient one, RR13 (TU4). In addition, RR13 repeated her comment two more times, in TU9 and TU15 at different stages of the rating process.

Excerpt 9.17: R4's and RR13's expressing uncertainty

R4 Script N4

TU	Rater talk
19	I don't even know...

RR13 Script N4

TU	Rater talk
4	I don't know

To sum up, the eleven subcategories in the coding scheme made the analysis of raters' comments possible. Own focus comments referred to text appearance, such as length, layout characteristics, eligibility and tidiness. Competent and proficient raters seemed to make similar comments on these surface features regarding both the number of occurrences and the content of comments. As far as those comments are concerned in which raters expressed their opinion, we can see that proficient raters made more comments in these subcategories which include reflection on script quality, overall impression or student proficiency. In addition, most raters verbalised their feelings during rating, the number of occurrences was the highest for all raters in this subcategory and proficient raters' comments outnumbered those of competent ones. Errors were frequently

corrected and proficient raters corrected mistakes more often than proficient ones, a strategy not promoted by either the scale or the training.

9.2 Conclusion

In this chapter I attempted to explore and analyse raters' foci during rating. Findings in Chapter Eight show that both competent and proficient raters attempted to follow the rating criteria suggested by the rating scale and this chapter provided a detailed analysis of their foci during the stages of the rating processes. Raters could internalise the rating sequence and they did not make many comments on the rating process. As far as their rating foci are concerned, they paid considerable attention to each of the four rating criteria; however, the most attention competent and proficient raters paid to was grammar. The differences are bigger when reading focus comments are investigated: although reading strategies occur frequently, competent raters did the most reading when rating task achievement. Task achievement was the criterion on which the two groups of raters behaved differently: proficient raters turned to reading strategies on fewer occasions than competent raters. Competent raters read more when they were dealing with grammar and their comments were less evenly distributed than those of proficient raters. Rating aspect of vocabulary attracted more attention from proficient raters than from competent ones and their distribution of comments was more even. Both groups read very little for rating the aspect of organisation. It seems that reading focus does not play a significant role in rating organisation or by then they know it all.

Looking into competent and proficient raters' own focus comments, I can conclude that in several cases, comment occurrences were similar in tendency, but the comments for the same script were different. Investigation into what criteria raters focused on apart from the suggested ones makes the observation of differences possible. Although both groups attended to similar surface features, proficient raters made more comments in some of the subcategories, such as length, eligibility and tidiness. It is also striking that proficient raters did not refrain from making evaluative comments and expressing their own feelings. In addition, they more often commented on students' proficiency and intentions.

The next chapter narrows the investigation further into rater behaviour and the four rating criteria are analysed in more detail to explore the similarities and differences between them and between competent and proficient raters as well.

Chapter 10

Raters' Focus on the Four Rating Criteria

Introduction

In the previous chapters I analysed the features of the rating patterns and raters' foci: I discussed raters' rating patterns in Chapter Eight from several aspects: I observed raters' gender distribution and language use in rating. Looking at language use in protocols I investigated length and sequencing features. Finally, I outlined the emerging rating patterns. Chapter Nine centred on raters' foci during rating: I compared competent and proficient raters' management, rating and reading strategies and investigated their own focus comments. This chapter attempts to look into further details of rating processes to consider raters' comments that can be explicitly related to any of the four rating criteria to answer the third research question:

3. How do competent and proficient raters interpret the four rating criteria: task achievement, vocabulary, grammar and organisation?

My discussion of these issues follows the sequence they appear in the rating scale without prioritising one over the other to reveal what criteria raters attended to when making rating decisions.

10.1 Raters' Focus on the Four Rating Criteria

Raters focused on different characteristics while rating the ten scripts. Their comments were grouped, as discussed in Chapter Nine, into four categories: management, rating, reading and own focus. Rating and reading focus comments were collected according to raters' explicit attention to the four rating criteria: task achievement, vocabulary, grammar and organisation.

Table 10.1
Number and Ratio of Raters' Foci (percentages in parentheses)

	Management (percentage)	Rating (percentage)	Reading (percentage)	Own focus (percentage)	Total (percentage)
Competent	1,027 (16%)	2,443 (38%)	2,053 (32%)	918 (14%)	6,441 (59%)
Proficient	746 (16%)	1,786 (40%)	1,043 (23%)	964 (21%)	4,539 (41%)
Total	1,773 (16%)	4,229 (39%)	3,096 (28%)	1,882 (17%)	10,980

Raters' focus on the four rating criteria, as Table 10.1 illustrates, represented the majority of their comments on script evaluation: there were 4,229 (39%) rating and 3,096 (28%) reading comments: 67% of all 10,980 remarks. What raters said with management and own focus was not included in the analysis, as these remarks refer to global points and could not be directly related to any of the four rating criteria. In addition, although such comments do play a role in rating processes, they are not explicitly directed towards any of the four points of task achievement, vocabulary, grammar and organisation: they include, for example, scoring criteria identification, speculations on students' proficiency, or error correction.

The analysis centres on strategies included in rating and reading categories to examine what competent and proficient raters attended to while evaluating the ten scripts. In addition, competent and proficient raters' comments are compared to understand the features of their rating processes. The following specific questions are investigated to find out

(a) whether raters paid equal attention to the scale descriptors,
(b) whether they used their own words for evaluation,
(c) what criteria not in the rating scale they applied,
(d) what additional strategies raters apply while rating.

These questions are in the centre of attention here and the four rating criteria are considered in turn. I investigated competent and proficient raters' rating and reading behaviour comments in the same manner when they were evaluating task achievement, vocabulary, grammar and organisation to allow comparisons.

10.2 Raters' Focus When Rating Task Achievement

The first rating aspect in the scale (see Appendix 7.2) was task achievement with two descriptors in each band to refer to the extent of achieving communicative goal and to the number of content points covered in compositions (see Figure 10.1 for details).

Rating aspect	Sub-category	Descriptors
Task achievement	Communicative goal	• Extent to which communicative goal is achieved
	Content points	• The number of content points covered

Figure 10.1. The rating aspect of task achievement in the rating scale

The rating aspect of task achievement focuses on the meaning of texts, as raters had to evaluate the extent to which writers of the letters achieved the communicative goal. In addition, raters had to decide how many content points students covered out of five included in the writing task (see the writing task in Appendix 7.1).

The coding scheme was developed so that raters' attention to the four rating criteria could be observed (see Appendix 7.8 for the complete coding scheme).

A Rating focus	Rating aspect	Code	Comment
	A1 Task achievement		
		A1a	Compares text to scale descriptor 1 communicative goal
		A1b	Compares text to scale descriptor 2 content points
		A1c	Evaluates the aspect in own words
		A1d	Adds own criterion
		A1e	Chooses score
		A1f	Adds reason why that score
		A1g	Revises decision
		A1h	Identifies error
		A1i	Refers to lack of detail
		A1j	Changes focus/switches to different criterion
		A1k	Finalises the score

Figure 10.2. Coding scheme extract for rating behaviour comments on task achievement

The extract from the coding scheme for rating behaviour comments on task achievement can be found in Figure 10.2.

As mentioned above, each of the four rating criteria was further divided into similar subcategories with the only difference in the first and second subcategories: codes A1a and A1b in the case of task achievement. Comment occurrences on each of the eleven subcategories by competent and proficient raters are illustrated in Figure 10.3. Competent and proficient raters attended to the different characteristics to a different extent: individual comment categories are analysed in detail below.

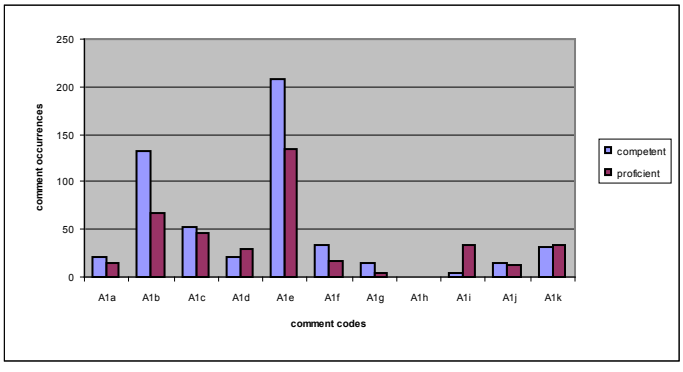


Figure 10.3. Rating criteria when rating task achievement

Apart from rating strategies, raters used different reading strategies in their rating processes when focusing on any of the four rating criteria. They include: reading the scale, the script and the task rubric for evaluation. In addition, raters gave examples and they sometimes summarised scripts, which is a new subcategory in the reading behaviour category of the coding scheme as the extract from the scheme in Figure 10.4 shows.

B Reading focus	Rating aspect	Code	Reading target
	B1 Task achievement		
		B1a	Scale
		B1b	Script: more words
		B1c	Summarises script
		B1d	Example: one word
		B1e	Rubric

Figure 10.4. Coding scheme extract for reading behaviour comments on task achievement

The distribution of comments competent and proficient raters made with a reading focus, as illustrated in Figure 10.5, shows that raters frequently turned to reading strategies and they read the scale, the script and the rubric most often.

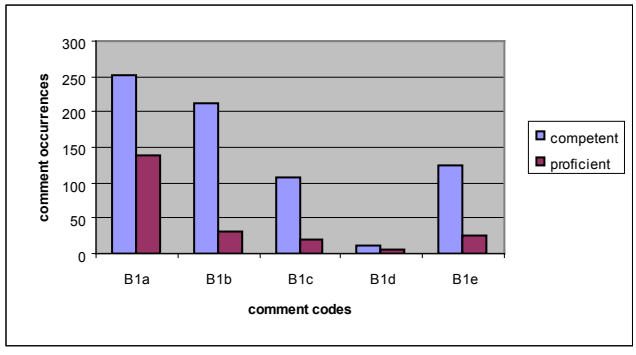


Figure 10.5. Reading behaviour when rating task achievement

Details of competent and proficient raters' attention to different reading strategies are discussed in the next sections.

There were two descriptors for rating task achievement in the rating scale: achievement of the communicative goal and the number of content points students covered. Occurrences of comments which reflect attention competent and proficient raters paid to evaluation of the communicative goal were similar: they made 21 (mean 1) and 14 (mean 1) comments, respectively.

Table 10.2
Attention to Scale Descriptor 1 When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1a	Compares text to scale descriptor 1: communicative goal	Competent		1	0	1	3	3	3	2	3	3	2	21 (1)	1.1
		Proficient		1	0	4	2	0	1	2	0	2	2	14 (1)	1.26
		Total		2	0	5	5	3	4	4	3	5	4	35 (.9)	

Table 10.2 illustrates that neither competent nor proficient raters talked about the achievement of the communicative goal of script N2 and proficient raters did mention communicative goal of scripts N5 and N8.

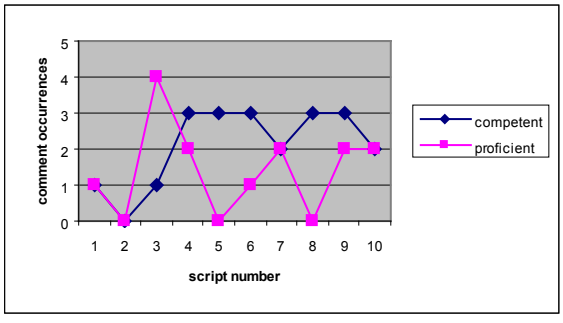


Figure 10.6. Comments on scale descriptor 1: communicative goal

The distribution curves in Figure 10.6 show the irregular feature of comments competent and proficient raters made on the achievement of communicative goal.

Raters referred considerably more often to the second descriptor than to the first, which was the number of content points covered. Competent raters made more remarks, 133 (mean 6) than proficient raters 68 (mean 4.5), on content points writers covered in their compositions (see Table 10.3 for details).

Table 10.3
Attention to Scale Descriptor 2: Content Points

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1b	Compares text to scale descriptor 2: content points	Competent		8	18	14	13	13	13	18	11	14	11	133 (6)	3.06
		Proficient		3	8	12	10	6	3	8	7	6	5	68 (4.5)	2.86
		Total		11	26	26	23	19	16	26	18	20	16	201 (5.4)	

The distribution of comments is similar in tendency, although competent raters dealt the most with scripts N2 and N7, and proficient raters with scripts N3 and N4, as comment occurrences show in Figure 10.7.

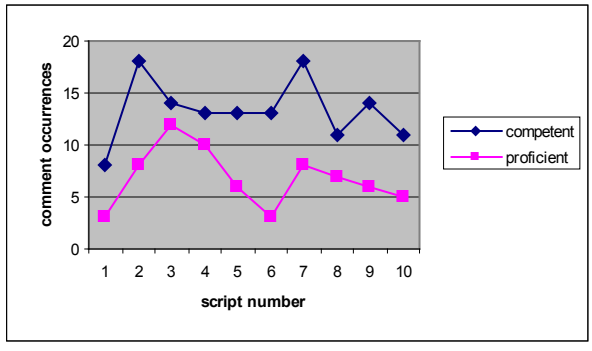


Figure 10.7. Comments on scale descriptor 2: content points

Competent and proficient raters turned to the rating scale and read out the descriptors, as Table 10.4 shows: almost all raters in both groups referred to the rating scale descriptors. Competent raters read 251 (mean 11.4) times and proficient raters fewer, 140 (mean 9.3) times. Looking at the comment distribution, we can see that scripts N3 and N8 attracted most often the competent raters' attention and script N7 the least.

Table 10.4
Reading the Scale Descriptors When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B1a	Reading the scale	Competent		24	21	34	28	28	21	17	28	25	25	251 (11.4)	4.77
		Proficient		18	9	15	11	16	18	13	11	13	16	140 (9.3)	3.09
		Total		42	30	49	39	44	39	30	39	38	41	391 (10.6)	

Proficient raters, on the other hand, read the most from the scale when rating script N1 and N6 and the least when rating script N2. Proficient raters' comments were more evenly distributed across the ten scripts, as it appears in Figure 10.8.

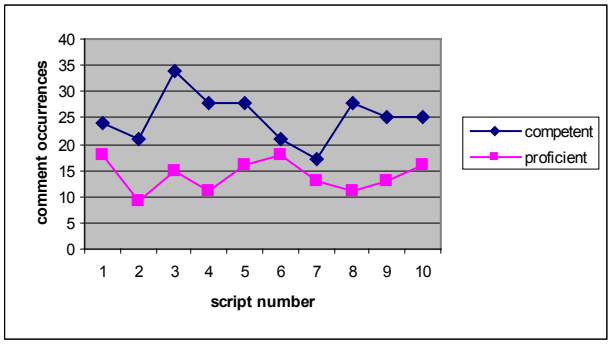


Figure 10.8. Reading the scale descriptors when rating task achievement

To sum up raters' reference to the rating scale, both competent and proficient raters were more engaged in evaluating content points than the achievement of the communicative goal. They read out scale descriptors more often than they compared the scripts to the scale descriptors. Competent raters made more comments when they were comparing scripts with the scale and they read out scale descriptors more frequently than proficient raters did.

The rating aspect of the task achievement was sometimes evaluated by raters using their own words, as Table 10.5 illustrates: competent raters evaluated scripts for task achievement less frequently in their own words than proficient raters.

Table 10.5
Script Evaluation in Own Words

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1c	Evaluates aspect in own words	Competent		9	3	1	5	14	6	3	7	1	4	53 (2.4)	3.97
		Proficient		5	1	5	5	5	6	3	7	4	5	46 (3.1)	1.65
		Total		14	4	6	10	19	12	6	14	5	9	99 (2.7)	

Competent raters used their own words 53 (mean 2.4) times, while proficient raters 46 (mean 3.1) times. The distribution of competent raters' comments in this subcategory was fairly uneven (sd 3.97) and proficient raters evaluated most scripts in own words similarly, except for script N2, which they commented once (see Figure 10.9).

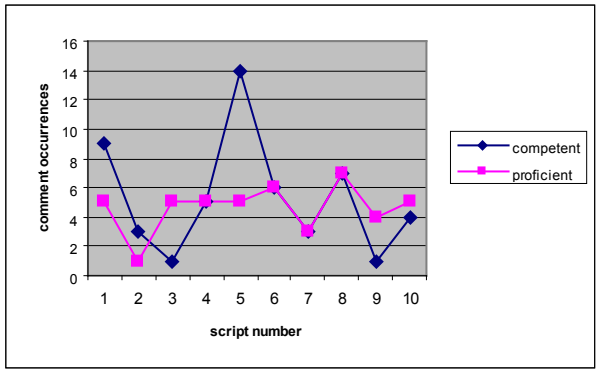


Figure 10.9. Script evaluation in own words

There were two scripts, N1 and N5, which generated the most of competent raters' wording and two scripts, N3 and N9 the fewest. Proficient raters' pattern was different: they paid the most attention to script N8 and the least to script N2.

When rating the aspect of task achievement, some raters added their criteria, competent raters contributed fewer times than proficient raters: 22 (mean 1) and 29 (mean 1.9) comments, respectively, as Table 10.6 shows.

Table 10.6
Additional Criteria When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1d	Adds own criterion	Competent		2	0	2	1	2	4	1	3	4	3	22 (1)	1.32
		Proficient		1	1	4	3	1	5	4	4	3	3	29 (1.9)	1.45
		Total		3	1	6	4	3	9	5	7	7	6	51 (1.4)	

The additional criteria not included in the rating scale were remarks on text variety and creativity. Raters frequently noticed organisational features when evaluating task achievement, they often referred to features, such as text coherence, cohesion or they said that proper letter characteristics were missing. They pointed at letter conventions evaluating task achievement, mentioned inappropriate salutation, lack of introduction and ending. There were also comments on the style of the letter.

The distribution of additional comments when rating task achievement was different for the two groups across the ten scripts, as Figure 10.10 demonstrates. Comparison of the two distribution curves shows that the number of remarks

for individual scripts was different; there were scripts, which competent raters attended to more than proficient ones (N1 and N9) and others were commented more often by proficient raters (e.g. script N7).

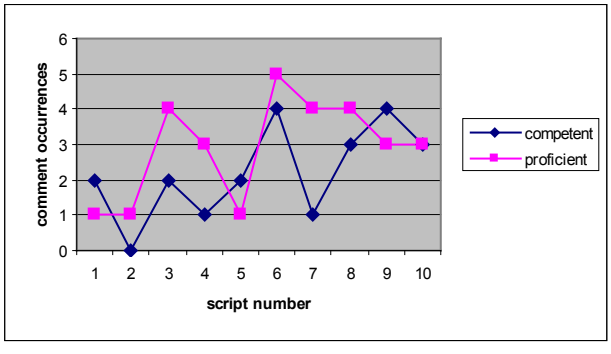


Figure 10.10. Additional criteria when rating task achievement

Both competent and proficient raters added their own criteria, although proficient raters referred to more additional features. These criteria mostly related to organisational features; some raters seemed to consider salutation, paragraphing and coherence to belong to the rating criterion of task achievement.

Evaluation of written performance ends in assigning scores on the basis of rating scales. Both competent and proficient raters duly announced their choice of scores; however, there were instances when score identification did not appear in the protocols. Table 10.7 shows the details of score identification: competent raters identified the score in most cases, 209 (mean 9.5) times and proficient raters 135 (mean 9) times.

Table 10.7
Score Nomination When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1e	Chooses score	Competent		22	20	21	21	21	20	20	21	22	21	209 (9.5)	.74
		Proficient		13	13	12	14	13	14	14	14	14	14	135 (9)	.71
		Total		35	33	33	35	34	34	34	35	36	35	344 (9.3)	

Even if some raters failed to announce the score they chose, all scores were duly entered in the score sheet they had to fill in while rating the scripts.

Rating processes of written performance are characterised by some strategies that raters turn to when rating scripts, they refer to the rating process and are closely related to the decision-making processes. These comments were grouped into six subcategories in each of the four rating criteria, as the extract from the coding scheme on rating task achievement in Figure 10.2 above shows: codes from A1f to A1k belong here.

Competent and proficient raters sometimes justified their decisions and they explicitly referred to the reason for the score. Comment occurrences in this subcategory are similar, as Table 13.8 illustrates: competent raters asserted their decision 34 (mean 1.5) times and proficient raters 17 (mean 1.1) times.

Table 10.8
Giving Reasons for Score Choice When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1f	Adds reason why that score	Competent		3	3	4	1	5	2	7	1	6	2	34 (1.5)	2.07
		Proficient		1	1	3	2	5	1	2	1	1	0	17 (1.1)	1.42
		Total		4	4	7	3	10	3	9	2	7	2	51 (1.4)	

The distribution of comment occurrences in Figure 10.11 shows that competent raters paid the most attention to scripts N7 and N9 (7 and 6 comments, respectively) and proficient raters to script N5 (5 comments).

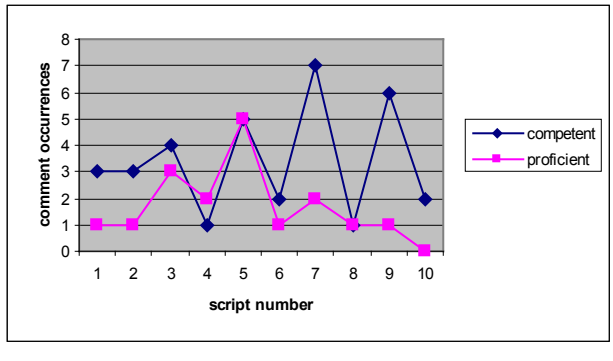


Figure 10.11. Giving reasons for score choice when rating task achievement

Competent raters provided justification less evenly than proficient raters, while none of the proficient raters justified their decision of script N10.

Competent and proficient raters did not often revise their decisions: there were a total 19 (mean .5) comments. Competent raters revised somewhat more than proficient raters (see Table 10.9). Competent raters revised their decision 14 (mean .6) times and proficient raters 5 (mean .3) times.

Table 10.9
Revision of Decision When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1g	Revises decision	Competent		1	3	1	1	2	0	4	0	2	0	14(.6)	1.35
		Proficient		1	0	1	1	0	0	0	1	1	0	5(.3)	.53
		Total		2	3	2	2	2	0	4	1	3	0	19(.5)	

There were two scripts, N6 and N10, whose scores were not revised by any of the raters; competent raters revised decision of seven scripts, while proficient raters revised scores of five scripts.

Raters did not reflect on errors when they were evaluating task achievement: they noticed errors twice altogether, as Table 10.10 shows, a competent rater on script N9 and a proficient rater on script N1.

Table 10.10
Error Identification When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total
A1h	Identifies and error	Competent		0	0	0	0	0	0	0	0	1	0	1
		Proficient		1	0	0	0	0	0	0	0	0	0	1
		Total		1	0	0	0	0	0	0	0	1	0	2

Competent and proficient raters remarked on lack of detail several times when they were rating task achievement. They referred to missing information 104 (mean 2.8) times (see Table 10.11).

Table 10.11
Comments on Lack of Detail When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1i	Refers to lack of detail	Competent		4	21	3	1	7	2	21	5	4	2	70 (3.2)	7.57
		Proficient		1	10	0	4	4	0	10	1	4	0	34 (2.3)	3.86
		Total		5	31	3	5	11	2	31	6	8	2	104 (2.8)	

Competent raters made more comments than proficient raters (70; mean 3.2 and 34; mean 2.3, respectively) and their comments were less evenly distributed (sd 7.57 and 3.86, respectively), which is illustrated in Figure 10.12.

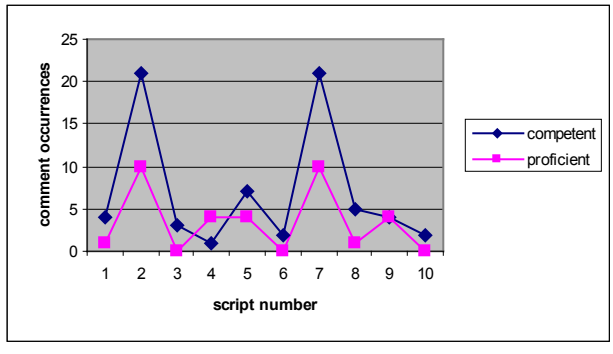


Figure 10.12. Comments on lack of detail when rating task achievement

The distribution of comments across the ten scripts shows similar tendencies, both competent and proficient raters found the most deficiencies in scripts N2 and N7, however, proficient raters referred to missing detail fewer times than competent raters.

As far as raters' changing focus is concerned, as Table 10.12 illustrates, they rarely changed their rating focus: a total of 26 (mean .7) comments were made.

Table 10.12
Change of Focus Comments When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1j	Changes focus switches to different criterion	Competent		4	3	2	2	1	1	0	0	1	0	14 (.6)	1.35
		Proficient		2	0	1	0	2	2	2	1	2	0	12 (.8)	.92
		Total		6	3	3	2	3	3	2	1	3	0	26 (.7)	

Competent raters changed focus somewhat less frequently than proficient raters: 14 (mean .6) versus 12 (mean .8).

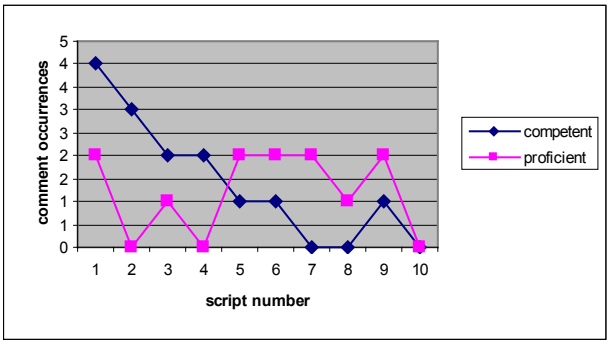


Figure 10.13. Change of focus when rating task achievement

The two distribution curves, as shown in Figure 10.3, illustrate different patterns: competent raters changed focus considerably more often when rating script N1 and their curve is descending. It implies that they could focus on the rating criterion more by the end of the rating process. Proficient raters' curve tendency shows a more even distribution, and their change of focus was less frequent, too.

Raters sometimes finalised the score they chose when they were rating task achievement, they did so 73 (mean 2) times.

Table 10.13
Score Finalisation When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A1k	Finalises score	Competent		3	4	4	3	7	2	8	3	2	3	39 (1.8)	2.02
		Proficient		4	7	4	4	2	2	3	0	5	3	34 (2.3)	1.9
		Total		7	11	8	7	9	4	11	3	7	6	73 (2)	

Table 13.13 shows that competent raters finalised scores 39 (mean 1.8) and proficient raters 34 (mean 2.3) times. According to the distribution curves across the ten scripts in Figure 10.14, competent raters finalised the score most often when they were rating scripts N5 and N7, while proficient raters when they were rating script N2.

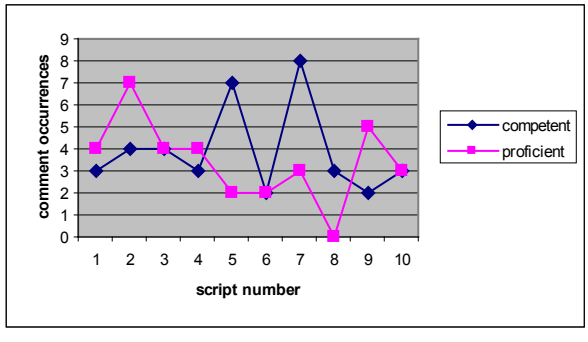


Figure 10.14. Score finalisation when rating task achievement

Raters' reading strategies when rating task achievement, as discussed above, included reading the scale, the script and the rubric. Raters' reading the scale was discussed above together with findings on raters' reference to the scale.

Competent and proficient raters either read more words from the scripts or cited individual words as examples. The two types of reading comments were grouped into two subcategories.

Table 10.14
Reading More Words from the Script When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B1b	Reading the script: more words	Competent		28	25	23	31	24	17	15	15	20	14	212 (9.6)	5.92
		Proficient		1	4	4	3	2	3	2	0	6	5	30 (2)	1.87
		Total		29	29	27	34	26	20	17	15	26	19	242 (6.5)	

Competent and proficient raters read texts extensively differently: competent raters 212 (mean 9.6) times, whereas proficient raters 30 (mean 2) times, which is a considerable difference. Competent and proficient raters' comments were distributed differently: standard deviation of competent raters' comments was 5.92, whereas for proficient raters it was 1.87 (see Table 10.14 for details).

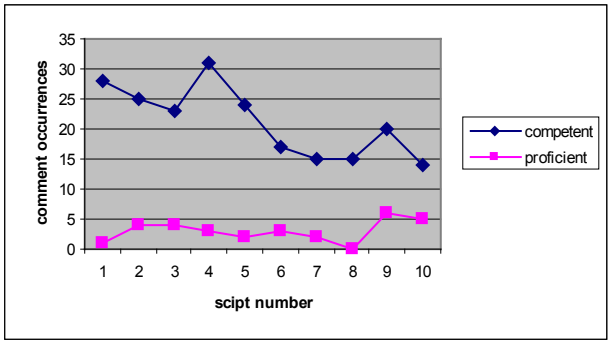


Figure 10.15. Reading more words from the script when rating task achievement

Apart from the large difference in comment occurrences, the distribution of comments across the ten scripts is also remarkable. Competent raters' curve shows a declining tendency, number of occurrences reduce considerably from 28 on script N1 to 14 on script N10, as Figure 10.15 illustrates, except for script N4 and N9. Proficient raters comments were more evenly distributed.

Table 10.15
Reading One-word Examples When Rating Task Achievement

Code	Comment type	Rater Script	1	2	3	4	5	6	7	8	9	10	Total (mean)
B1d	Reading an example: one word	Competent	1	4	0	2	0	1	1	1	0	0	10(.5)
		Proficient	0	0	3	0	0	1	0	1	0	0	5(.3)
		Total	1	4	3	2	0	2	1	2	0	0	15(.4)

Reading one word as an example did not occur frequently, as is shown in Table 10.15, there were a total of 15 (mean .4) examples cited: competent raters quoted more (10; mean .5) than proficient raters (5; mean .3).

Competent raters summarised script contents more often than proficient raters: competent raters 107 (mean 4.9) times, while proficient raters on 20 (mean 1.3) occasions, which is a considerable difference (see Table 10.16). In addition, competent raters turned to this strategy variously across the scripts, they summarised most often when rating script N5 (18 comments) and least often when rating script N8 (6 comments). The content of scripts N1 and N4 was not summarised by any of the proficient raters.

Table 10.16
Summarising Scripts When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B1c	Summarises the script	Competent		12	14	7	10	18	9	11	6	10	10	107 (4.9)	3.43
		Proficient		0	5	2	0	3	1	3	2	2	2	20 (1.3)	1.49
		Total		12	19	9	10	21	10	14	8	12	12	127 (3.4)	

Proficient raters' summaries were more evenly distributed (see Figure 13.16) than those of competent raters, though there was a script, N2, which received more attention than the others.

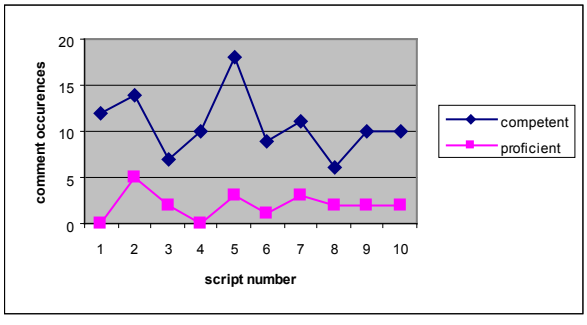


Figure 10.16. Summarising the scripts when rating task achievement

Reading the rubric was the last reading comment type in the coding scheme. The difference between competent and proficient raters' comment occurrences shows that competent raters referred considerably more to this strategy than proficient raters. There were a total of 151 (mean 4.1) references to rubric, as Table 10.17 illustrates: competent raters read rubric 125 (mean 5.9) times, whereas proficient raters 26 (mean 1.7) times.

Table 10.17
Reading the Rubric When Rating Task Achievement

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B1e	Reading the rubric	Competent		11	19	8	8	12	11	18	16	13	9	125(5.7)	3.98
		Proficient		5	1	0	2	3	0	4	3	4	4	26(1.7)	1.78
		Total		16	20	8	10	15	11	22	19	17	13	151(4.1)	

Distribution of references to rubric across the ten scripts, as it appears in Figure 10.17 shows similar pattern for the two groups of raters, though competent raters paid the most attention to scripts N2 and N7 (19 and 18 comments, respectively), and proficient raters to script N1 (5 comments).

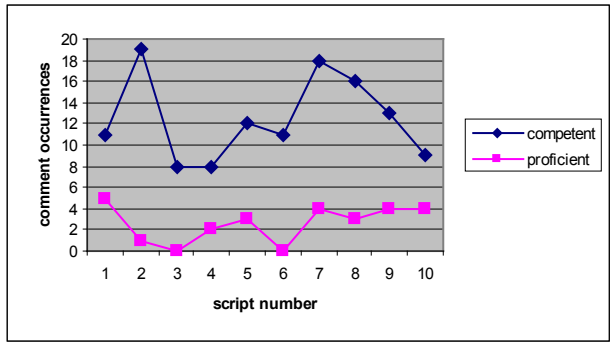


Figure 10.17. Reading the rubric when rating task achievement

When rating task achievement competent and proficient raters paid attention to the different rating criteria: investigation into comment occurrences revealed differences not only between the two rater groups, but in comment distribution among the ten scripts. The most remarkable difference was in the reading behaviour between the two groups of raters: competent raters referred to reading strategies more often than proficient raters. Raters' focus when rating task achievement along the lines of the specific research questions can be summarised as follows:

a. Findings on raters' focus when they were rating task achievement show that they did not pay equal attention to both scale descriptors: they seemed to be more engaged in evaluating content points than achievement of communicative goal. However, they were reading the scale descriptors, especially competent raters turned to this strategy frequently.
b. Raters sometimes evaluated the rating aspect of task achievement in their own words; proficient raters did so more often than competent raters.
c. When rating task achievement raters did not use many of their own criteria, there were some occurrences and if they verbalised their own criteria in rating, these were similar and referred to creativity, variety or organisational features of the scripts. In addition, raters often changed rating focus and referred to one of the other three rating criteria (vocabulary, grammar or organisation) when dealing with task achievement.
d. The occurrences of comments on reading behaviour were considerably higher for competent raters than for proficient ones. Rating task achievement involved missing detail identification, reading the task rubric and summarising content.

Finally, there are some additional observations on differences in raters' attention:

a. Occurrences of competent raters' comments that reflect their changing rating focus show decreasing tendency which can mean that they internal-

ised rating criteria over time and could differentiate between them more by the end of the rating process.

b. A decrease in comments could be observed when competent raters were reading out more words from scripts. This trend may indicate that they did not need to refer to reading strategies by the end of their rating processes as extensively as at the beginning, as they were already familiar with the texts.

10.3 Raters' Focus When Rating Vocabulary

The second rating criterion in the rating scale was vocabulary with one descriptor, which included two rating criteria: range and appropriacy (see Figure 10.18 for details).

Rating aspect	Sub-category	Descriptor
Vocabulary	Range	Range of words and expressions
	Appropriacy	Appropriacy of words and expressions

Figure 10.18. The rating aspect of vocabulary in the rating scale

Rating the aspect of vocabulary involved evaluation of variety of words and expressions writers used and raters had to judge if the extent of writers' choice of vocabulary items was appropriate.

Similarly to the other three rating criteria, the coding scheme included several subcategories, as the extract from the rating behaviour section of the coding scheme in Figure 13.19 demonstrates (the complete coding scheme can be found in Appendix 7.8).

A Rating focus	Rating aspect	Code	Strategy
	A2 Vocabulary		
		A2a	Compares text to scale descriptor 1: range
		A2b	Compares text to scale descriptor 2: appropriacy
		A2c	Evaluates the aspect in own words
		A2d	Adds own criterion
		A2e	Chooses score
		A2f	Adds reason why that score
		A2g	Revises decision
		A2h	Identifies error
		A2i	Refers to lack of detail
		A2j	Changes focus/switches to different criterion
		A2k	Finalises the score

Figure 10.19. Coding scheme extract for rating related comments on vocabulary

The coding scheme aided the observation of the rating behaviour for evaluation of vocabulary and included eleven subcategories, the first two of which, A2a and A2b referred to specific scale descriptors of the rating criterion of vocabulary and the other nine were identical with the subcategories for rating criteria of task achievement, grammar and organisation.

Rating the scripts on the four rating criteria involved not only rating strategies, but reading strategies as well. Raters read the scale, the scripts and the rubric when they were rating the four criteria, including the aspect of vocabulary. In addition, they sometimes summarised script content. Reading behaviour comments in the coding scheme were the same for the four rating criteria; however, each of them had its own code indicating which of the four aspects it relates to, as an extract for rating vocabulary in Figure 10.20 illustrates.

B Reading focus	Rating aspect	Code	Reading target
	B2 Vocabulary		
		B2a	Scale
		B2b	Script: more words
		B2c	Summarises script
		B2d	Example: one word
		B2e	Rubric

Figure 10.20. Coding scheme extract for reading related comments on vocabulary

Comment occurrences of competent and proficient raters on the eleven subcategories when rating the aspect of vocabulary, as Figure 10.21 illustrates showed different patterns.

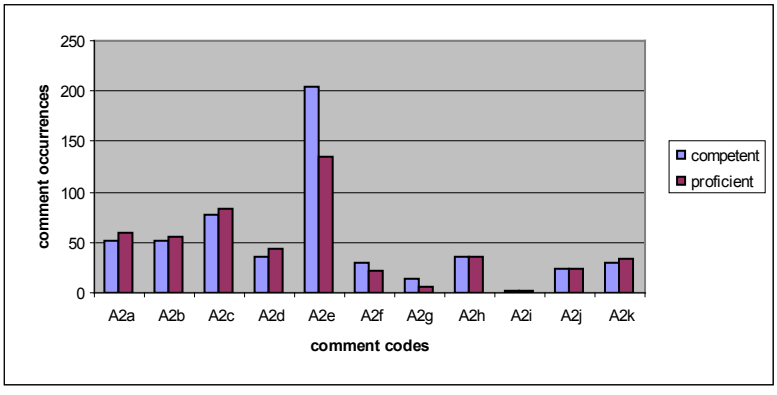


Figure 10.21. Rating criteria of vocabulary evaluation

Raters reflected on each of the eleven comment categories, most often they talked about score nomination and the least about lack of detail. Competent and proficient raters attended to the scale descriptors frequently; however, they preferred to evaluate the aspect in their own words. They added their criteria, justified scores, identified errors, changed focus and finalised the scores similarly, and they revised their decisions as well.

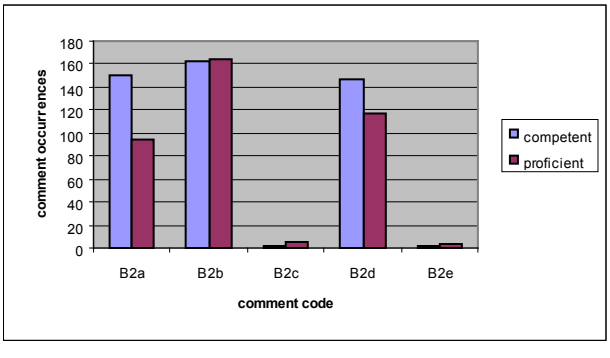


Figure 10.22. Reading comments for vocabulary

Raters' reading focus, as Figure 10.22 shows, was very different: raters paid the most attention to reading the scale, more words and one-word examples from the scripts. Neither competent nor proficient raters turned to the strategies

of summarising content and reading the rubric very often. Rating and reading comment occurrences are analysed in more detail in the following sections.

Next, I examine raters' attention to the rating scale by looking into the occurrences of rating and reading behaviour comments. Raters' reference to the vocabulary range, as it appears in Table 10.18, was different for competent raters, who compared texts to the scale considerably fewer times (51; mean 2.1) than proficient ones (59; mean 4).

Table 10.18
Comparing Text to Scale Descriptor 1: Range When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2a	Compares to scale descriptor 1 range	Competent		6	4	1	7	4	7	6	5	8	3	51 (2.1)	2.13
		Proficient		6	3	6	7	5	7	4	5	7	9	59 (4)	1.73
		Total		12	7	7	14	9	14	10	10	15	12	110 (3)	

Proficient raters' remarks were more evenly distributed across the scripts: the standard deviation figure of proficient raters' comments was lower (1.73) than that of competent raters (2.13).

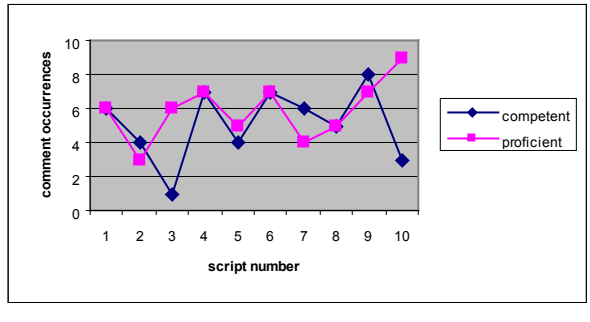


Figure 10.23. Comments on scale descriptor 1: range when rating vocabulary

The distribution curves show a variety in the occurrence patterns, the biggest differences were when raters evaluated scripts N3 and N10; proficient raters said more about vocabulary range (6 and 9, respectively), whereas in the case of scripts N2, N7 and N9 competent raters talked more, as illustrated in Figure 10.23.

The rating criterion of appropriacy attracted raters' attention similarly to the criterion of range: there were a total of 107 (mean 2.9) comments (see Table

10.24 for details). Likewise, the criterion of vocabulary range, competent raters talked less about accuracy (51 comments; mean 2.3) than proficient raters, who made 56 (mean 3.7) remarks in this subcategory.

Table 10.19
Comparing Text to Scale Descriptor 2: Appropriacy When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2b	Compares to scale descriptor appropriacy	Competent		9	4	1	6	5	5	2	6	10	3	51 (2.3)	2.85
		Proficient		8	6	3	7	3	6	5	8	8	2	56 (3.7)	2.27
		Total		17	10	4	13	8	11	7	14	18	5	107 (2.9)	

The distribution across the ten scripts shows that competent raters' comments were less evenly distributed than those of proficient raters.

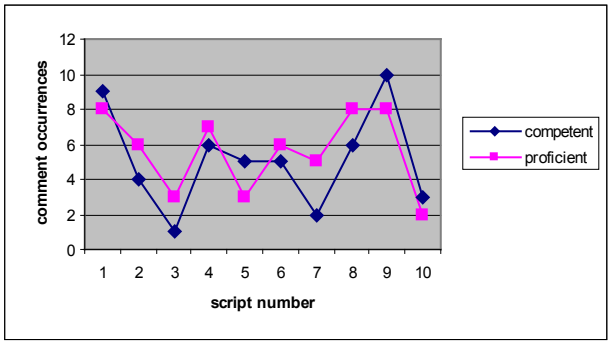


Figure 10.24. Comments on scale descriptor 2: appropriacy when rating vocabulary

However, the two curves are closer in pattern, but competent raters paid the most attention to scripts N1 and N9 and the least to scripts N3 and N7 as it appears in Figure 10.24. Proficient raters talked more when evaluating scripts N1, N8 and N9 and less when dealing with script N10.

The rating scale played a considerable role in evaluation, as raters referred to reading the scale very frequently: they read the scale 271 (mean 7.3) times (see Table 10.20). Competent raters read more 177 (mean 8), while proficient raters referred to reading the scale less often, as there were 94 (mean 6.3) reading comments in the subcategory of reading the scale.

Table 10.20
Reading the Scale Descriptors When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B2a	Reading the scale	Competent		19	16	27	25	20	12	19	14	13	12	177 (8)	5.29
		Proficient		12	10	10	12	6	6	8	13	9	8	94 (6.3)	2.46
		Total		31	26	37	37	26	18	27	27	22	20	271 (7.3)	

Proficient raters' remarks were more evenly distributed, as standard deviation figures (5.29 and 2.46, respectively) and the comment occurrence curves in Figure 10.25 illustrate.

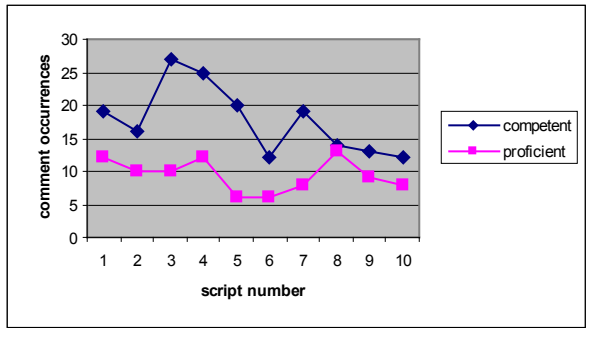


Figure 10.25. Reading the scale when rating vocabulary

The two distribution curves show some similarity as far as their patterns are concerned; however, competent raters read the scale the most often when rating script N3 and proficient raters when they were dealing with script N8.

To sum up, raters paid considerable attention to scale descriptors, which is reflected in their comment occurrences. Their focus on rating and reading strategies in rating vocabulary was similar, although competent raters used somewhat more reading than rating strategies, whereas proficient raters read the scale on fewer occasions.

Table 10.21
Evaluation in Own Words When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2c	Evaluates the aspect with own words	Competent		11	9	9	7	8	8	6	5	5	10	78(3.5)	2.04
		Proficient		7	8	10	3	11	9	13	9	7	7	84(5.6)	2.72
		Total		18	17	19	10	19	17	19	14	12	17	162(4.4)	

Raters evaluated the aspect of vocabulary in their own words 84 (mean 4.4) times, as Table 10.21 indicates. Competent raters used own words on 78 (mean 3.5) occasions, while proficient raters considerably more frequently (84; mean 5.6).

The two curves of comment occurrence distribution are very different for competent and proficient raters (see Figure 10.26).

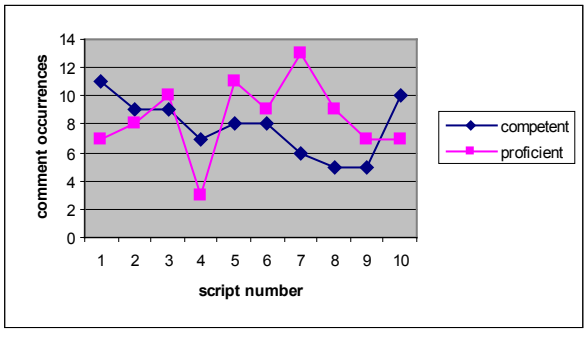


Figure 10.26. Evaluation in own words when rating vocabulary

Competent raters' comments are more evenly distributed than those of proficient raters, and the pattern of attention they paid to individual scripts is also different. Competent raters evaluated in most words scripts N1 and N10 (11 and 10 comments, respectively) and scripts N8 and N9 using much less language (5 comments each). Proficient raters evaluated script N7 with the most words and script N4 with the fewest (13 and 3 comments, respectively).

Table 10.22
Adding Own Criterion When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2d	Adds own criterion	Competent		0	3	3	3	7	4	5	4	3	3	35 (1.6)	1.78
		Proficient		1	9	4	4	7	3	4	5	4	2	43 (2.9)	2.31
		Total		1	12	7	7	14	7	9	9	7	5	78 (2.1)	

Raters sometimes added their own criteria to the rating aspect of vocabulary: there were 78 (mean 2.1) remarks in the raters protocols that contained criteria not included in the scale (see Table 10.22).

Although these criteria were similar in kind, competent raters referred less often to them than proficient raters: competent raters made 35 (mean 1.6) comments and proficient raters 43 (mean 2.9). The majority of raters in their additional criteria referred to spelling mistakes and inaccuracies. They sometimes pointed out word repetition, lifting words from the rubric or referred to word number in scripts. Competent raters' comments across the ten scripts were distributed more evenly than proficient raters' remarks (see Figure 10.27).

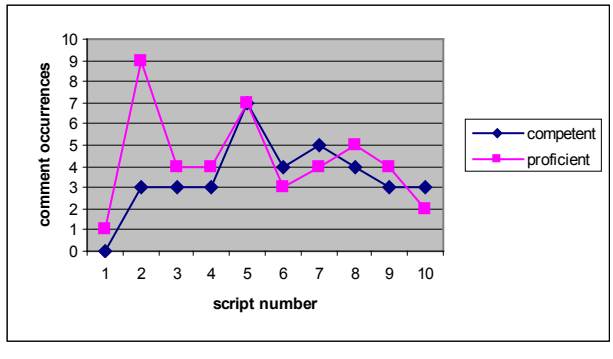


Figure 10.27. Adding own criterion when rating vocabulary

Competent raters added the most criteria when rating script N5 (7) and they did not report any on script N1. Proficient raters mentioned the most additional criteria when rating script N2 and N5 (9 and 7 comments, respectively) and they did not come up with any for script N1. The two patterns show some similarities: the attention paid to scripts N1 and N5 was similar for the two groups of raters.

The rating process includes a stage in which raters are expected to choose a score to quantify their evaluation.

Table 10.23
Choosing a Score When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2e	Chooses score	Competent		19	21	21	21	21	20	21	21	19	21	205 (9.3)	.85
		Proficient		15	13	13	13	12	14	14	11	14	15	134 (9)	1.26
		Total		34	34	34	34	33	34	35	32	33	36	339 (9.2)	

Raters chose a score and almost always announced it, as it appears in Table 10.23, there were 339 (mean 9.2) occurrences of score choice. Competent raters announced the score more often than proficient raters, there were 205 (mean 9.3) comments and 134 (mean 9) made respectively; however, raters entered all scores in the score sheets.

Additional strategies that raters turned to when rating the aspect of vocabulary, similarly to the rating criteria of task achievement, grammar and organisation, included adding a reason for the score, indication of decision of revision, error identification, reference to lack of detail, changing focus and score finalisation. However, competent and proficient raters used these strategies differently not only within one rating criterion, but for each of the four aspects. The intention is to explore the different strategies raters turned to when rating vocabulary in the following sections.

Competent and proficient raters added reason for the score making 51 (mean 1.4) comments in this subcategory. According to the data in Table 10.24, the number of comment occurrences is similar, however they are distributed across the ten scripts differently, as Figure 10.29 illustrates.

Table 10.24
Adding Reason for the Score When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2f	Adds reason why that score	Competent		4	3	4	1	3	4	2	1	3	5	30 (1.4)	1.33
		Proficient		4	4	1	1	2	1	2	1	3	2	21 (1.4)	1.2
		Total		8	7	5	2	5	5	4	2	6	7	51 (1.4)	

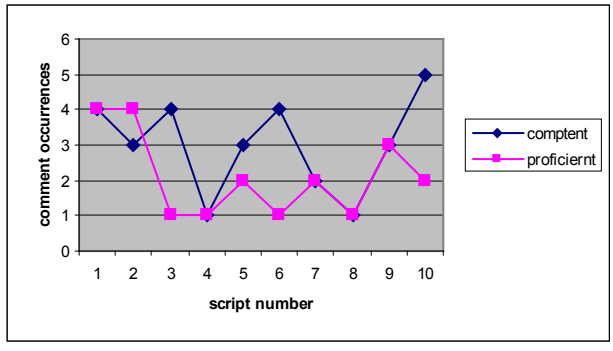


Figure 10.28. Adding reason for the score when rating vocabulary

The two curves are different, competent raters justified the score most often in case of script N10 (5 comments), while proficient raters in case of scripts N1 and N2 (4 comments each). The least attention in this subcategory was paid to scripts N4 and N8 by competent raters and to scripts N3, N4, N6 and N8 by proficient raters, which was one comment for each script.

Raters rarely revised their decisions when rating vocabulary, as Table 10.25 shows, there were 20 (mean .5) comments made altogether, none of the raters revised their decision when evaluating script N10.

Table 10.25
Revision of Decisions When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)
A2g	Revises decision	Competent		0	1	3	0	1	2	2	3	2	0	14 (.6)
		Proficient		1	2	0	1	0	0	0	2	0	0	6 (.4)
		Total		1	3	3	1	1	2	2	5	2	0	20 (.5)

Competent raters referred to this strategy more often than proficient raters, there were 14 (mean .6) and 6 (mean .4) comments, respectively in this subcategory.

Competent and proficient raters rarely identified errors when they were rating vocabulary: 36 (mean 1.9) times altogether (see Table 10.26 for details).

Table 10.26
Error Identification When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2h	Identifies an error	Competent		4	4	2	4	4	1	0	3	10	4	36 (1.6)	2.67
		Proficient		9	3	2	3	3	1	3	1	11	0	36 (2.4)	3.57
		Total		13	7	4	7	7	2	3	4	21	4	72 (1.9)	

Competent raters identified errors on fewer occasions, there were 36 (mean 1.6) remarks, while proficient raters, who identified errors in 36 (mean 2.4) comments and they were more unevenly distributed.

The two distribution curves, as they appear in Figure 10.29, are not identical; still, there are some similarities in patterns: for example, script N9 attracted the most attention regarding error identification, competent and proficient raters pointed at errors to a similar extent (10 and 11 comments, respectively). The most significant difference was in error identification for script N1: competent raters reflected on errors less frequently (4 comments) than proficient (9 comments).

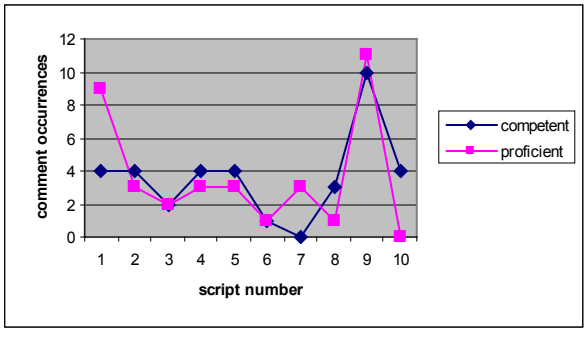


Figure 10.29. Error identification when rating vocabulary

Raters did not notice missing details in scripts regarding the rating aspect of vocabulary; there were three scripts, as Table 10.27 illustrates, in which lack of detail was detected: a competent rater mentioned a missing detail in script, N6 and two scripts N5 and N8 were mentioned by proficient raters with a comment each.

Table 10.27
Comments on Lack of Detail When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total
A2i	Refers to lack of detail	Competent		0	0	0	0	0	1	0	0	0	0	1
		Proficient		0	0	0	0	1	0	0	1	0	0	2
		Total		0	0	0	0	1	1	0	1	0	0	3

Raters changed their focus and attended to another criterion when rating vocabulary 48 (mean 1.3) times, competent raters made fewer such comments (total 24, mean 1.1) and they were more evenly distributed (sd 1.35) than comments made by proficient raters (total 24, mean 1.6) as is shown in Table 10.28.

Table 10.28
Changing Focus When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2j	Changes focus/ switches on different criterion	Competent		3	2	2	2	0	2	2	2	5	4	24 (1.1)	1.35
		Proficient		3	7	1	2	2	0	0	2	4	3	24 (1.6)	2.07
		Total		6	9	3	4	2	2	2	4	9	7	48 (1.3)	

Most raters changed focus to grammar and there were some raters who changed to evaluation of organisation or task achievement while they were rating vocabulary.

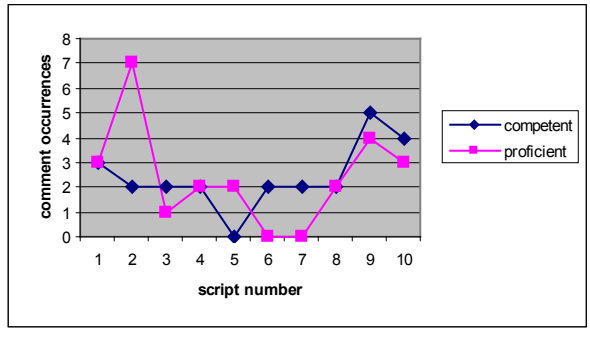


Figure 10.30. Changing focus when rating vocabulary

The comment distribution curves are similar in some respect, as Figure 13.30 shows, but script N2 received more comments (7) in this subcategory from competent raters than from proficient ones (2).

Raters rarely confirmed a score: there were 63 (mean 1.7) such comments made, as it appears in Table 10.29, competent raters referred to finalisation less often (29; mean 1.3) than proficient raters (34; mean 2.3).

Table 10.29
Finalising the Score When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A2k	Finalises score	Competent		3	3	3	4	3	6	1	4	1	1	29 (1.3)	1.6
		Proficient		5	4	3	4	5	2	4	4	2	1	34 (2.3)	1.35
		Total		8	7	6	8	8	8	5	8	3	2	63 (1.7)	

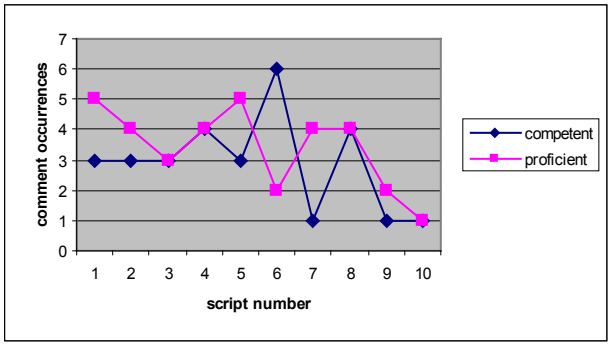


Figure 10.31. Finalising the score when rating vocabulary

Proficient raters' comment distribution is more even than that of competent raters; however, as in Figure 10.31, the two curves are very different. Competent raters attended the most to script N6 (6 comments) and the least to scripts N9 and N10 (a comment for each), whereas proficient raters finalised score most often for scripts N1 and N5 (5 comments for each), and the fewest times for script N10.

Raters' reading strategies were divided into two groups: one group included comments in which raters read more words from the scripts, and the one-word examples were placed in the second group.

There were 328 (mean 8.9) occasions when raters read more words from the scripts. As Table 10.30 shows, competent raters read from scripts less often than proficient raters. Competent raters read out more words 163 (mean 7.4) than

proficient raters 165 (mean 11) times and their comments were more evenly distributed (sd 7.07) than proficient raters (sd 7.66).

Table 10.30
Reading More Words from the Scripts When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B2b	Reading the script: more words	Competent		19	21	15	20	10	15	12	3	29	19	163 (7.4)	7.07
		Proficient		17	25	14	10	7	27	13	6	23	23	165 (11)	7.66
		Total		36	46	29	30	17	42	25	9	52	42	328 (8.9)	

The occurrence patterns are similar in tendency, as shown in Figure 10.32, except for script N6, which received the most comments (27) from proficient raters and much fewer (15) from competent raters.

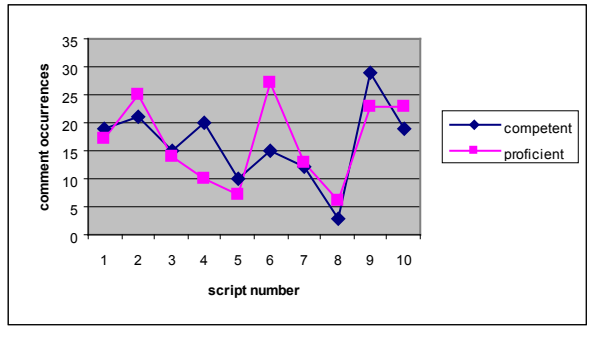


Figure 10.32. Reading more words from the scripts when rating vocabulary

Raters read out one-word examples less frequently than longer texts, there were 117 (mean 7.1) such comments. There is some difference between the occurrences: it was lower for competent (146; mean 6.6) than for proficient raters (117; mean 7.8). Proficient raters' comments were more evenly distributed; the standard deviation is 5.81 for proficient raters and 7.5 for competent raters (see Table 10.31).

Table 10.31
Reading One Word Examples When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B2d	Reads example: one word	Competent		17	24	11	10	11	9	21	10	28	5	146 (6.6)	7.5
		Proficient		8	18	1	18	10	7	20	11	11	13	117 (7.8)	5.81
		Total		25	42	12	28	21	16	41	21	39	18	263 (7.1)	

The distribution of comments across the ten scripts is similar in tendency in some cases, for example scripts N1, N2 and N5; however, there is a big difference in connection with script N9, as competent raters read out 28 one-word examples, whereas proficient raters read 11 words (see Figure 10.33).

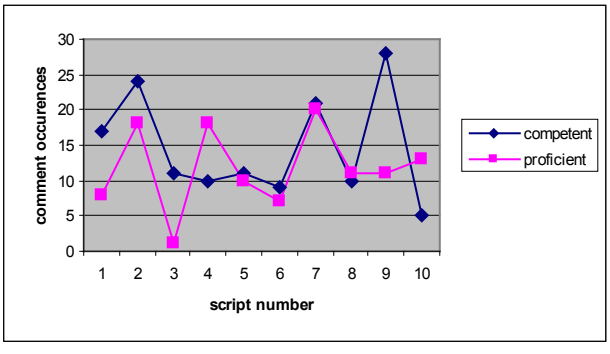


Figure 10.33. Reading one-word examples when rating vocabulary

Raters summarised the scripts on seven occasions when they were rating vocabulary. Two competent raters summarised script N2 and five proficient raters did so when rating four scripts, as Table 10.32 shows.

Table 10.32
Summarising Script When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)
B2c	Summarises script	Competent		0	2	0	0	0	0	0	0	0	0	2
		Proficient		0	0	0	0	1	2	1	1	0	0	5
		Total		0	2	0	0	1	2	1	1	0	0	7

Competent and proficient raters did not refer to the rubric when rating vocabulary, except for five occasions, as Table 10.33 illustrates. The comment distribution pattern is very different, competent raters read the rubric when they were dealing with scripts N1 and N2, while proficient raters made three comments for script N8.

Table 10.33
Reading the Rubric When Rating Vocabulary

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total
B2e	Reads the rubric	Competent		1	1	0	0	0	0	0	0	0	0	2
		Proficient		0	0	0	0	0	0	0	3	0	0	3
		Total		1	1	0	0	0	0	0	3	0	0	5

The comparisons of competent and proficient raters' foci revealed similarities and differences in rating behaviour. The main observations according to the specific research questions are as follows:

a. Competent and proficient raters paid similar attention to both scale descriptors when rating vocabulary; however, proficient raters used somewhat more rating comments than competent raters. On the other hand, competent raters read the rating scale more often.
b. Raters often evaluated the rating criterion of vocabulary in their own words, especially proficient raters turned to this strategy frequently.
c. Occurrences of own criteria were not high, although proficient raters added more criteria, but they were similar for the two groups and they mainly referred to spelling errors and inaccuracies. In some cases raters did not seem to be able to distinguish criteria related to vocabulary from criteria related to grammar. When raters changed focus in rating vocabulary, it was for focus on grammar.
d. Raters paid considerable attention to the scale, they either rated the scripts referring to the scale descriptors, or they read out the scale descriptors. In addition, they frequently evaluated the criterion in their own words. As far as the scripts are concerned, raters read the scripts many times, as is indicated by high occurrences of reading related comments. Raters did not turn to the rubrics when they rated vocabulary.

The analysis of competent and proficient raters' vocabulary assessment is followed by the next rating aspect in the rating scale, which is grammar.

10.4 Raters' Focus When Rating Grammar

The third rating criterion in the rating scale was grammar with two descriptors in each band (see Appendix 7.2 for the complete rating scale). One descriptor was for evaluating accuracy, and the other one to assess structures in the texts (see Figure 10.34 for details).

Rating aspect	Subcategory	Descriptor
Grammar	Accuracy	Occurrence and feature of inaccuracies
	Structures	Variety of structures

Figure 10.34. The rating aspect of grammar in the rating scale

Raters had to decide to what extent inaccuracies in the texts hindered comprehension and what range of structures writers used to convey the message in their letter.

Similarly to task achievement, vocabulary and organisation, for observation of rater behaviour a coding scheme was employed according to which raters' comments were categorised. As mentioned above, the coding scheme categories were identical regarding the aspects of task achievement, vocabulary, grammar and organisation except for the first two subcategories which referred to scale descriptors explicitly. The extract from the coding scheme in Figure 10.35 illustrates the rating focus in the coding scheme (the complete coding scheme is in Appendix 7.8). The eleven subcategories are identified with a code and the first two are directly related to the rating aspect of grammar, i.e. they are for rating accuracy and variety of structures. Apart from rating strategies, raters used reading strategies in rating and they are analysed in the same way as rating comments.

A Rating focus	Rating aspect	Code	Comment
	A3 Grammar		
		A3a	Compares text to scale descriptor 1: accuracy
		A3b	Compares text to scale descriptor 2: structures
		A3c	Evaluates the aspect in own words
		A3d	Adds own criterion
		A3e	Chooses score
		A3f	Adds reason why that score
		A3g	Revises decision
		A3h	Identifies error
		A3i	Refers to lack of detail
		A3j	Changes focus/switches to different criterion
		A3k	Finalises the score

Figure 10.35. Coding scheme extract for rating comments on grammar

These categories in the coding scheme were identical for the four rating criteria except for the codes, which were specific to the rating aspect. The extract for the reading behaviour comments in the rating scale can be found in Figure 10.36.

B Rating focus	Rating aspect	Code	Reading target
	B 3 Grammar		
		B3a	Scale
		B3b	Script: more words
		B3c	Summarises script
		B3d	Example: one word
		B3e	Rubric

Figure 10.36. Coding scheme extract for reading comments when rating grammar

Competent and proficient raters' rating processes are traced according to these categories in the following sections to get an insight into raters' decision-making behaviour.

Competent and proficient raters evaluated grammar referring to each of the eleven rating behaviour subcategories, as it appears in Figure 10.37.

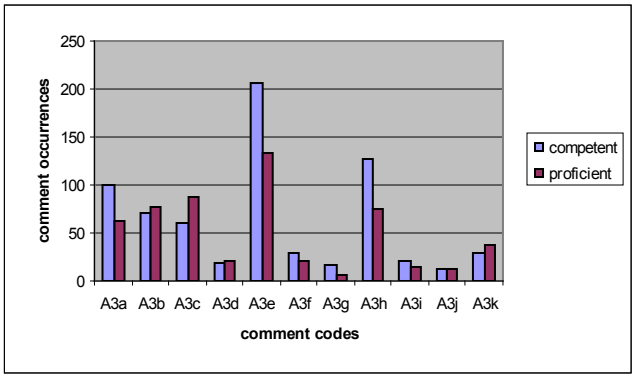


Figure 10.37. Rating criteria of grammar evaluation

As comment occurrences reveal, they paid the most attention to score nomination, error recognition and the rating criteria. There was a distinction between reference to rating the criteria of grammar: whether raters referred to the rating scale (codes A3a and A3b) or evaluated the criterion in their own words (code A3c). There were considerably more comments in these subcategories than in those, which included comments that raters' made during rating to add a criterion not in the rating scale (A3d), justify the score (A3f), revise decision (A3g), refer to lack of detail (A3i), change focus (A3j), or finalise the score (A3k). There were variations not only in comment use, but there were differences between the two groups of raters. In the following sections a detailed analysis is provided of different features of raters' rating behaviour.

Apart from the rating strategies, raters turned to reading strategies during which competent and proficient raters referred to the scale, the script and the rubric.

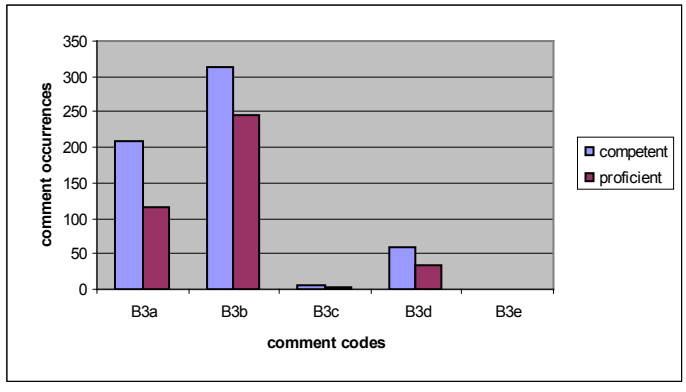


Figure 10.38. Reading comments when rating grammar

As Figure 10.38 illustrates, neither competent nor proficient raters referred to the rubric (code B3e) when they were evaluating the rating aspect of grammar.

Raters read extensively from texts (code B3b) and they read from the scale (code B3a) more than they read one-word examples (code B3d). There were few occasions when raters summarised the texts (code B3c). Comment occurrences are discussed in the following sections and competent and proficient raters' reading behaviour is compared and an investigation of comment occurrences is carried out.

Competent and proficient raters' focus on scale descriptors can be observed by looking into the occurrences of comments on the two scale descriptors and reading out scale descriptors.

First, I examine competent and proficient raters' attention to the first scale descriptor: accuracy. As Table 10.34 shows, competent and proficient raters referred to accuracy similarly; however, competent raters made considerably more, 100 (mean 4.5) comments than proficient raters (total 63, mean 4.2) and their comments were more evenly distributed across the ten scripts.

Table 10.34
Comparing Scripts to Scale Descriptor 1: Accuracy

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3a	Compares text to scale descriptor 1: accuracy	Competent		8	11	10	12	14	12	5	11	9	8	100 (4.5)	2.58
		Proficient		6	5	7	11	4	9	6	7	3	5	63 (4.2)	2.36
		Total		14	16	17	23	18	21	11	18	12	13	163 (4.4)	

The two comment occurrence distribution curves, as they appear in Figure 10.39, show different tendencies: competent raters paid the most attention to script N5 (14 comments), while proficient raters' four comments were the second fewest in this subcategory for that script.

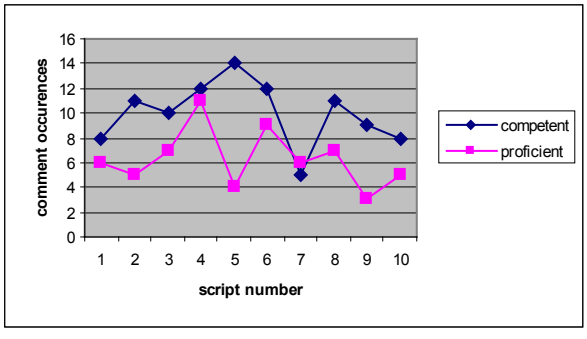


Figure 10.39. Comparing scripts to scale descriptor 1: accuracy

Raters focused on the criterion of variety of structures similarly to the criterion of accuracy, there were 147 (mean 4) contributions in this respect, however, as Table 10.35 indicates, competent raters made 70 (mean 3.2) comments, whereas proficient raters made more, 77 (mean 5.1) on structures and they were more evenly distributed (sd 2.31) than competent raters' comments (sd 2.36).

Table 10.35
Comparing Scripts to Scale Descriptor 2: Structures

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3b	Compares text to scale descriptor 2: structures	Competent		5	6	9	4	5	7	6	8	12	8	70 (3.2)	2.36
		Proficient		5	9	8	6	9	4	8	7	12	9	77 (5.1)	2.31
		Total		10	15	17	10	14	11	14	15	24	17	147 (4)	

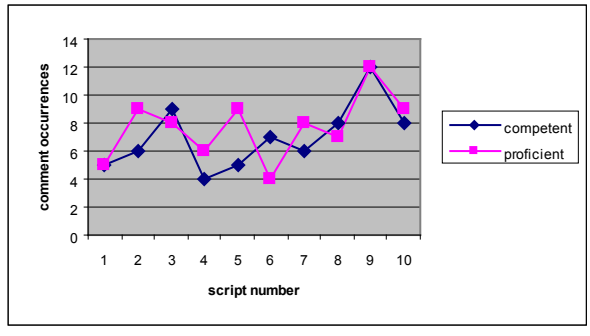


Figure 10.40. Comparing scripts to scale descriptor 2: structures

The distribution curves are similar in tendency across the ten scripts (see Figure 13.40); for example, more attention was paid to scripts N3 and N9 than to script N4 by both groups of raters.

Raters turned to the rating scale and they often read the scale descriptors: 326 (mean 8.8) times (see Table 10.36 for details). Competent raters read out from the scale more often (210; mean 9.5) than proficient raters (116; mean 7.7) and proficient raters' comments were more evenly distributed than those of competent raters (sd 3.17 and 3.86, respectively).

Table 10.36
Reading the Scale When Evaluating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B3a	Reads the scale	Competent		22	20	22	21	30	19	17	20	23	16	210 (9.5)	3.86
		Proficient		16	16	13	14	9	8	11	10	7	12	116 (7.7)	3.17
		Total		38	36	35	35	39	27	28	30	30	28	326 (8.8)	

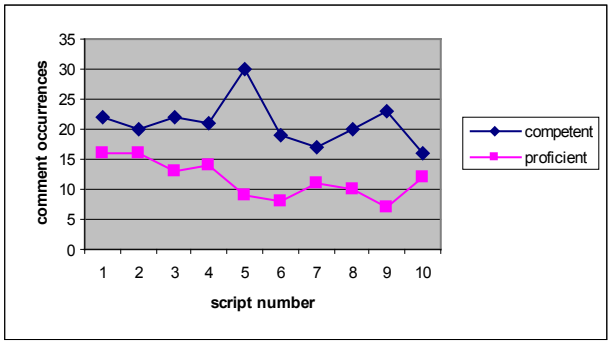


Figure 10.41. Comments on reading the scale when rating grammar

The patterns of occurrences are different: competent raters read the most from the scale when they were dealing with scripts N5 and N9, while proficient raters read the most when rating scripts N1 and N2. Proficient raters' comment distribution curve shows a declining tendency: there were more comments made at the beginning than at the end of the rating process.

Competent and proficient raters frequently evaluated grammar using their own words; they did so 149 (mean 4) times, as demonstrated in Table 10.40. Proficient raters' evaluation in own words occurred considerably more times: there were 88 (mean 5.9) remarks they were more evenly distributed (sd 2.25) than the comments made by competent raters. Competent raters' 61 (mean 2.8) comments were less evenly distributed (sd 2.6).

Table 10.37
Raters' Evaluation of Grammar in Own Words

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3c	Evaluates aspect in own words	Competent		6	2	3	8	8	9	7	6	3	9	61 (2.8)	2.6
		Proficient		11	12	9	10	8	10	7	6	5	10	88 (5.9)	2.25
		Total		17	14	12	18	16	19	14	12	8	19	149(4)	

The distribution curves in Figure 13.44 are very different: competent raters evaluated scripts N6 and N9 (9 comments each) with most own worded comments, while proficient raters paid the most attention to script N2 (12 comments) which competent raters commented the least (2 comments).

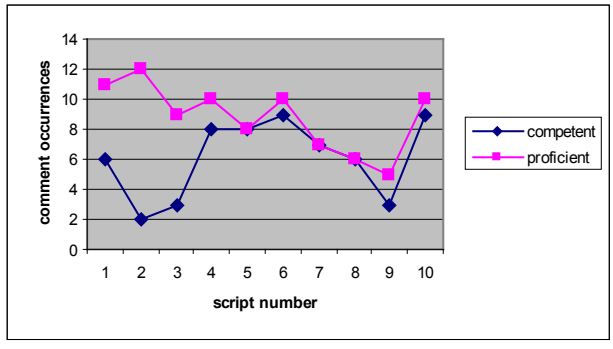


Figure 10.42. Evaluation of grammar in own words

Raters' did not add many criteria when they were rating grammar: there were altogether 40 (mean 1.1) occurrences of such remarks. Competent and proficient raters added criteria differently, as Table 10.41 demonstrates, competent raters made somewhat fewer comments (19; mean .8) than proficient raters (21; mean 1.4). However, the additional criteria were similar: out of the total of 40 comments 18 (45%) were remarks on word order by competent and proficient raters alike. They also added the criterion of punctuation, article use, repetition, style and complexity of sentences when rating grammar.

Table 10.38
Additional Criteria When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3d	Adds own criterion	Competent		3	0	1	7	7	0	0	0	1	0	19 (.8)	2.85
		Proficient		2	3	3	1	2	1	2	5	0	2	21 (1.4)	1.37
		Total		5	3	4	8	9	1	2	5	1	2	40 (1.1)	

The comments were unevenly distributed across the ten scripts, as Figure 10.43 shows. Competent raters added the most of own criteria when rating scripts N4 and N5 (7 comments each) and they did not add any criteria for scripts N2, N6, N7, N8 and N10.

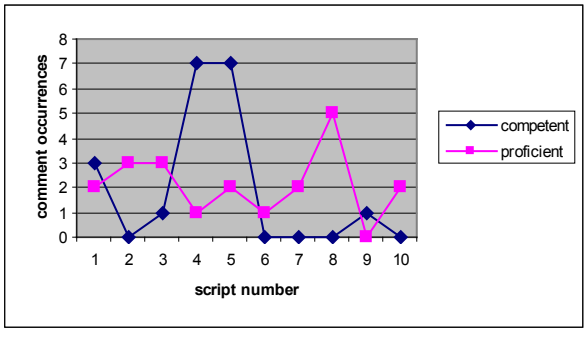


Figure 10.43. Additional criteria when rating grammar

On the other hand, proficient raters added their own criteria to all but one (N9) scripts. They paid the most attention to script N8.

As Table 10.39 shows, almost all raters announced their choice of the score: on 340 (mean 9.2) occasions. Competent raters verbalised their choice more often than proficient raters (206 times; mean 9.4), who said what the score was 134 (mean 8.9) times. However, all scores were entered in the score sheet in the rating pack, even if they did not appear in the think-aloud protocols.

Table 10.39
Score Nomination When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3e	Chooses score	Competent		20	18	22	22	20	20	20	21	22	21	206 (9.4)	1.26
		Proficient		13	12	13	14	13	13	16	12	14	14	134 (8.9)	1.17
		Total		33	30	35	36	33	33	36	33	36	35	340 (9.2)	

The comment distribution curves for both groups are similar in tendency, as Figure 10.46 shows, in case of script N7, there was one more comment observed than the number of raters in the proficient group. This happened because the rater most probably changed his/her mind and announced the new score as well.

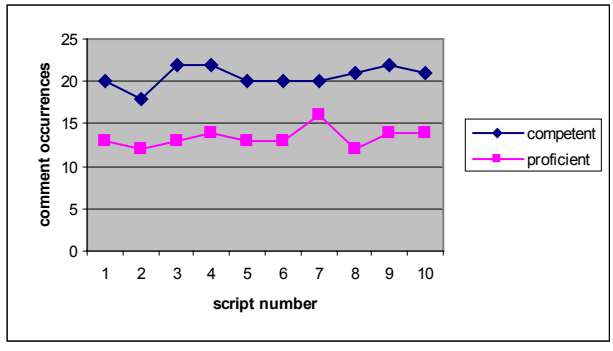


Figure 10.44. Score nomination when rating grammar

As mentioned earlier, raters turned to some additional strategies in evaluation of the rating aspect, which aided their decision-making processes. These strategies are investigated in detail in the following sections.

Competent and proficient raters rarely turned to the strategy of score justification, there were altogether 49 (mean 1.3) comments (see Table 10.43 for details). Their attention was similar: competent raters added a reason 29 (mean 1.3) times and proficient raters 20 (mean 1.3) times, but the distribution of comments across the ten scripts was different.

Table 10.40
Adding Reason Why That Score When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3f	Adds reason why that score	Competent		3	3	6	1	0	3	6	1	5	1	29 (1.3)	2.18
		Proficient		3	0	2	1	3	1	2	2	4	2	20 (1.3)	1.15
		Total		6	3	8	2	3	4	8	3	9	3	49 (1.3)	

The two curves differ, as Figure 10.45 shows, especially competent raters' comment distribution seems extreme: they paid the most attention to scripts N3 and N7 with 6 remarks on each, and did not justify their score at all in case of script N5. Proficient raters, on the other hand, justified their scores for all scripts, except N2 and made one to four justifications on each script.

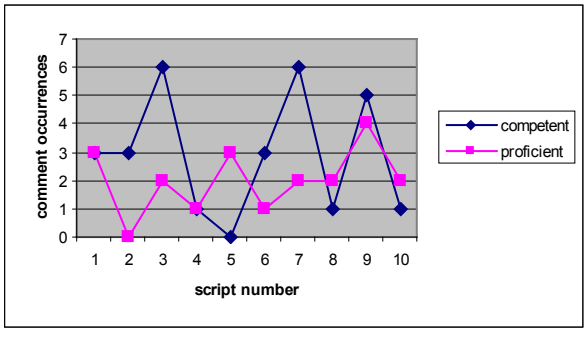


Figure 10.45. Adding reason why that score when rating grammar

Raters did not revise their decisions frequently, they made 22 (mean 1.5) comments, as Table 10.41 illustrates, there was a script, N1 which none of the raters commented on. Competent raters revised their decisions more (16; mean .7) than proficient raters (6; mean .4).

Table 10.41
Revision of Decisions When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3g	Revises decision	Competent		0	2	3	3	2	2	2	0	1	1	16 (.7)	1.07
		Proficient		0	0	0	0	0	2	2	1	0	1	6 (.4)	.84
		Total		0	2	3	3	2	4	4	1	1	2	22 (1.5)	

Apart from script N1, competent raters did not revise their decisions when rating the aspect of grammar for script N8. Proficient raters, on the other hand, revised their decisions only when they were dealing with scripts N6, N7, N8 and N9 with one or two comments.

Error identification often occurred when rating grammar, as can be seen in Table 10.42, raters identified 203 (mean 5.5) errors. Competent raters noticed more, 128 (mean 5.8) than proficient raters, who mentioned 75 (mean 5) errors.

Table 10.42
Error Identification When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3h	Identifies error	Competent		11	11	17	22	13	8	14	11	12	9	128 (5.8)	4.1
		Proficient		14	3	16	7	9	2	8	7	6	3	75 (5)	4.6
		Total		25	14	33	29	22	10	22	18	18	12	203 (5.5)	

The distribution of comment occurrences (see Figure 10.46 for details) is similar with some exceptions. Competent raters identified the most errors when they were rating script N4 (22), while proficient raters' noticed fewer errors (7) in that script. There was one script, N1 which attracted more proficient raters' attention than that of competent ones.

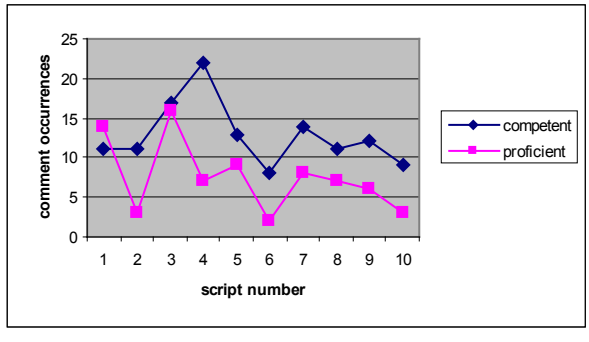


Figure 10.46. Error identification when rating grammar

Competent and proficient raters referred to a lack of detail to the same extent: there were 35 (mean 1) comments made altogether: 21 (mean 1) and 14 (mean 1) remarks, respectively. The distribution of comments is uneven; there was a script, N8, which was not commented on and scripts N9 and N10 received the most attention by all raters (see Figure 10.47 for details).

Table 10.43
Reference on Lack of Detail When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3i	Refers to lack of detail	Competent		1	2	1	1	1	1	1	0	7	6	21 (1)	2.38
		Proficient		1	0	3	2	0	0	1	0	4	3	14 (1)	1.51
		Total		2	2	4	3	1	1	2	0	11	9	35 (1)	

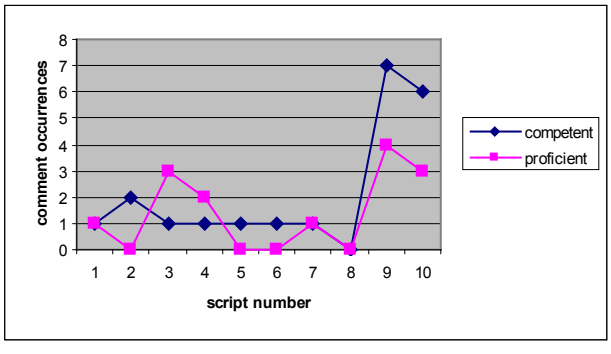


Figure 10.47. Comments on lack of detail when rating grammar

In addition, proficient raters did not identify a lack of detail in scripts N2, N5 N6 and N7.

Raters did not change focus often when rating grammar. There were altogether 25 (mean .7) occasions when they did so, as it appears in Table 10.44. Proficient raters changed focus less often, 13 (mean .6) times than proficient raters, who made 12 (.8) comments which referred to different focus from the rating criterion of grammar. There were two scripts, N5 and N6 which were not commented on in this category at all.

Table 10.44
Changing Focus When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3j	Changes focus/ switches to another criterion	Competent		1	3	1	1	0	0	2	2	2	1	13 (.6)	.9
		Proficient		1	3	1	3	0	0	1	1	2	0	12 (.8)	1.14
		Total		2	6	2	4	0	0	3	3	4	1	25 (.7)	

Although there were differences in comment occurrences between the two groups of raters, they were similar for both groups. Most raters changed focus from grammar to vocabulary: there were 17 comments (66%) out of a total of 25. Some raters changed focus for task achievement, organisation and there was one rater who was talking about comprehensibility of text while rating the aspect of grammar.

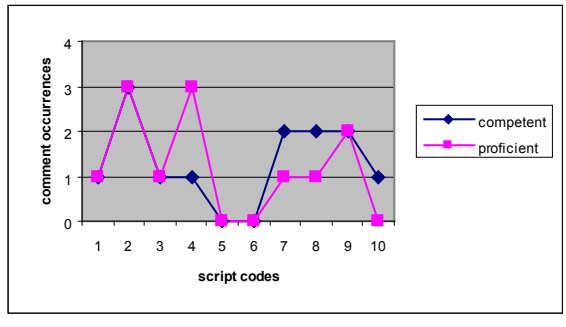


Figure 10.48. Changing focus when rating grammar

The comment distribution curves are different, as Figure 10.48 shows, especially in case of script N4 which was commented on once by a competent and three times by proficient raters.

Competent and proficient raters sometimes finalised scores: 66 (mean 1.8) times, as Table 10.48 illustrates. Competent raters finalised their score for grammar on fewer occasions than proficient raters. Competent raters did so 29 (mean 1.3) times and proficient raters 37 (mean 2.5) times, respectively.

Table 10.45
Finalising the Score When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A3k	Finalises the score	Competent		2	7	3	1	2	3	3	1	4	3	29 (1.3)	1.73
		Proficient		1	3	5	3	4	5	5	5	5	1	37 (2.5)	1.64
		Total		3	10	8	4	6	8	8	6	9	4	66 (1.8)	

The distribution of comments across the scripts, as it appears in Figure 10.49 is different for competent and proficient raters: there were some scripts which generated more comments from competent raters (e.g. N1 and N2), and there were scripts which proficient raters mentioned more often (e.g. N6 and N7).

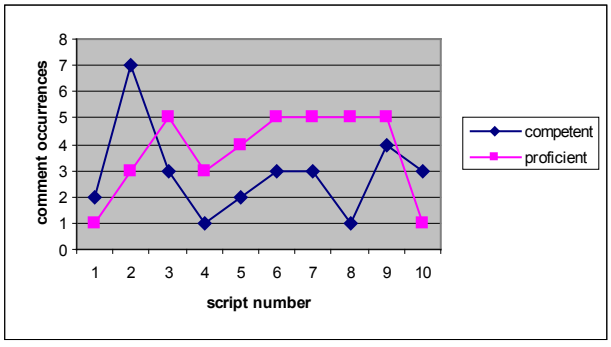


Figure 10.49. Finalising the score when rating grammar

Reading played a role in rating the scripts; however, as mentioned above, there were no comments on reading the rubric when raters evaluated grammar. I discussed reading the scale with rating focus above and raters reading and summarising the texts are analysed here. Two types of reading were identified: (1) raters read more words from the texts and (2) they cited one word only.

The occurrences of reading more words from the scripts are collected in Table 10.46: raters read texts extensively 560 times (mean 15.1), competent raters read more from scripts, 314 (mean 14.3) parts, while proficient raters 246 (mean 16.4) parts and proficient raters' comments were more evenly distributed (sd 7.49) than those of competent raters (sd 11.17).

Table 10.46
Reading the Script: More Words When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B3b	Reads the script: more words	Competent		25	16	30	50	46	22	38	38	27	22	314 (14.3)	11.17
		Proficient		23	15	32	40	20	22	29	18	27	20	246 (16.4)	7.49
		Total		48	31	62	90	66	44	67	56	54	42	560 (15.1)	

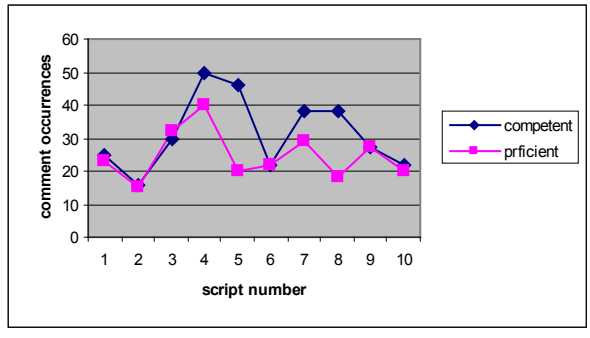


Figure 10.50. Reading the script: more words when rating grammar

The two curves of score distribution across the ten scripts are similar, as illustrated in Figure 10.50, there were two scripts, where there was considerable difference between the raters' attention: script N5 and N8. Competent raters read more when dealing with these scripts (46 and 38 comments, respectively) than proficient raters (20 and 18 comments, respectively).

Raters sometimes read out one-word examples: 94 (mean 2.5) words and their distribution was considerably uneven (sd 4.01 and sd 2.22, respectively), especially for competent raters (see Table 10.47 for details). Competent raters read more words, 59 (mean 2.7) than proficient raters 35 (mean 2.3).

Table 10.47
Reading One Word as an Example When Rating Grammar

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B3d	Reads the script: one word	Competent		1	4	11	10	4	7	1	12	6	3	59 (2.7)	4.01
		Proficient		5	1	4	8	4	1	5	3	3	1	35 (2.3)	2.22
		Total		6	5	15	18	8	8	6	15	9	4	94 (2.5)	

The distribution of comments across the ten scripts shows extreme occurrences, mainly competent raters' comments had a wide range, they read out one example from scripts N1 and N7 and twelve words from script N8 (see Figure 10.53 for details). Proficient raters cited a word from scripts N2, N6 and N10 and the most words from script N4.

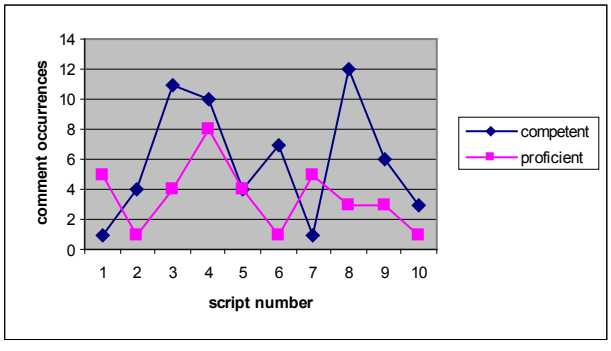


Figure 10.51. Reading one word as an example when rating grammar

Raters rarely turned to summarising the text content when rating the aspect of grammar; there were eight comments altogether, as is shown in Table 10.48. On five scripts they summarised content once or twice: N1, N3, N6, N9 and N10 either competent or proficient raters or both.

Table 10.48
Summarising Scripts When Rating Grammar

Code	Comment type	Rater / Script	1	2	3	4	5	6	7	8	9	10	Total
B3c	Summarises script	Competent	1	0	2	0	0	1	0	0	1	1	6
		Proficient	1	0	0	0	0	0	0	0	0	1	2
		Total	2	0	2	0	0	1	0	0	1	2	8

Competent and proficient raters paid careful attention to the rating criterion of grammar, there were comments on each of the eleven rating and five reading subcategories except for reading of the rubric. The patterns of distribution of comments across the ten scripts were considerably different for both competent and proficient raters, which can lead to a tentative conclusion that evaluation of the rating criterion of grammar depends on the quality of texts; some texts generate more comments of the same kind than others do.

Looking at the specific research questions, the following observations were made:

(e) Raters paid attention to the scale descriptors differently: competent raters seemed to attend more to accuracy than to structures, but overall they made fewer comments on comparing the scripts to the scale descriptors than proficient raters did. As far as reading the scale descriptors is concerned, competent raters read the scale descriptors more often than proficient raters. To sum up, competent raters preferred reading the scale descriptors to com-

paring them to texts, while proficient raters did the opposite: they read the scale fewer times than they compared them to the scripts.

(f) As for raters using their own words for evaluation of the aspect, we could see that proficient raters turned to this strategy more often than competent raters did.

(g) The additional criteria competent and proficient raters used for rating grammar were similar in kind; they added the criterion of word order the most often. They included some other criteria, such as punctuation, article use, word repetition and style. Competent and proficient raters justified their score and indicated lack of detail similarly, but made comments on different scripts. Competent raters revised their decisions and identified errors more often than proficient raters did. They frequently switched focus to the rating criterion of vocabulary when evaluating grammar, which can lead to a tentative conclusion that in some cases raters merged rating grammar and vocabulary. More proficient raters finalised their scores on the rating criterion of grammar than competent raters.

(h) Findings show that raters paid considerable attention to the rating scale and scripts, but all of them ignored the rubric entirely when dealing with the rating criterion of grammar.

The fourth rating criterion in the rating scale was organisation, which is discussed in the sections that follow.

10.5 Raters' Focus When Rating Organisation

The fourth criterion in the rating scale was organisation with two descriptors in each band. The two descriptors referred to layout features and to coherence of texts, as Figure 10.52 shows.

Rating aspect	Subcategory	Descriptor
Organisation	Layout features	How much the script reminds the reader of a letter
	Text coherence	Logical link between the elements of the text

Figure 10.52. The rating aspect of organisation in the rating scale

Raters had to pay attention to the extent to which each script resembled a letter and they were expected to evaluate links between the elements of texts (The detailed presentation of the rating scale is provided in Chapter Ten and the complete rating scale can be found in Appendix 7.2).

The observation of raters' rating processes similarly to rating aspects of task achievement, vocabulary and grammar was carried out using a coding scheme which aided categorisation of raters' comments. The extract of the rating aspect

for organisation from the coding scheme in Figure 10.53 illustrates the codes for rating the aspect of organisation (the complete coding scheme can be found in Appendix 7.8).

A Rating focus	Rating aspect	Code	Comment
	A4 Organisation		
		A4a	Compares text to scale descriptor 1: layout
		A4b	Compares text to scale descriptor 2: links
		A4c	Evaluates the aspect in own words
		A4d	Adds own criterion
		A4e	Chooses score
		A4f	Adds reason why that score
		A4g	Revises decision
		A4h	Identifies error
		A4i	Refers to lack of detail
		A4j	Changes focus/switches to different criterion
		A4k	Finalises the score

Figure 10.53. Coding scheme extract for rating comments on organisation

The coding scheme subcategories were identical for the four aspects of task achievement, vocabulary, grammar and organisation, except for the first two codes (A4a and A4b), which were specific to the rating criterion: the first subcategory includes comments in which raters compared the script to the first scale descriptor on layout. The second subcategory comprises comments related to the links between text elements.

Reading strategies were also categorised and the extract from the coding scheme is shown in Figure 10.54. The five subcategories for reading behaviour comments were identical for the four criteria; the only difference was in codes, which were specific for the rating aspect of organisation.

B Reading focus	Rating aspect	Code	Reading target
	B4 Organisation		
		B4a	Scale
		B4b	Script: more words
		B4c	Summarises script
		B4d	Example: one word
		B4e	Rubric

Figure 10.54. Coding scheme extract for reading comments on organisation

Competent and proficient raters' attention to organisation, as Figure 10.55 illustrates, was different.

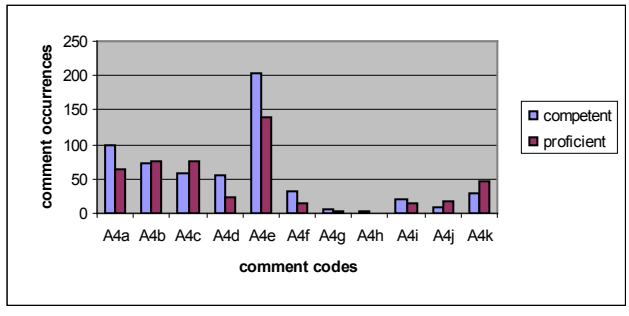


Figure 10.55. Raters' comments when rating organisation

Raters focused on the rating criterion and compared scripts to the scale descriptors or evaluated the rating aspect in their own words. They frequently nominated the score they chose. Those comments that were not explicitly related to the rating criterion mostly included some additional criteria, score justification and score finalisation. Raters sometimes revised their decisions, identified errors, referred to a lack of detail and changed focus as well. A detailed analysis and comparison of competent and proficient raters' rating behaviour is provided in the following sections.

Raters rarely turned to reading strategies, as comment occurrences in Figure 10.56 show; they mainly read the scale when rating the aspect of organisation. They did not read either more words or one-word examples from the scripts frequently, and they hardly ever summarised the scripts. There were only three comments by proficient raters on the rubric and none by competent raters.

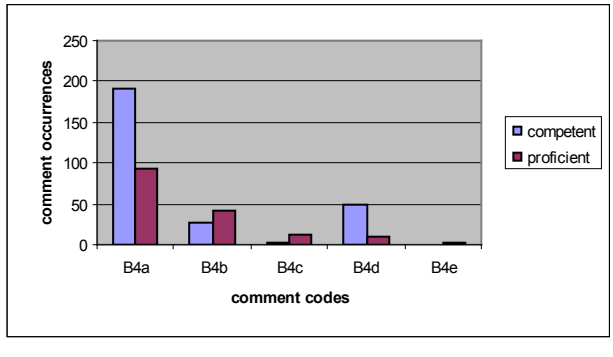


Figure 10.56. Reading related comments when rating organisation

I present raters' reading-related comments in more depth and compare competent and proficient raters' reading strategies in the following sections.

Raters' focus on scale descriptors can be observed by looking into comment occurrences in three subcategories in the coding scheme: when raters compared texts to scale descriptor on the layout (A4a) and to the descriptor on the links in the text (A4b). In addition, raters read out the scale descriptors (B4a) while they were rating the scripts.

First, I discuss how competent and proficient raters evaluated the layout of the scripts: they made 163 (mean 4.4) comments when they compared the scripts to the first descriptor in the scale, as Table 10.49 shows.

Table 10.49
Comparing Scripts to Scale Descriptor 1: Layout When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4a	Compares text to scale descriptor 1: layout	Competent		9	8	9	14	14	11	7	7	9	10	98 (4.5)	2.53
		Proficient		7	6	7	5	9	6	6	6	8	5	65 (4.3)	1.27
		Total		16	14	16	19	23	17	13	13	17	15	163 (4.4)	

Competent raters expressed their opinion more often (98; mean 4.5) than proficient raters, who made 65 (mean 4.3) comments. Competent raters' comments were less evenly distributed (sd 2.53) than those of proficient raters (sd 1.27). The two distribution curves in Figure 10.57 demonstrate how remarks were distributed across the ten scripts. Competent raters paid the most attention to scripts N4 and N5 and proficient raters to script N5 and N9. The least attention competent raters paid to scripts N7 and N8 and proficient raters to scripts N4 and N10.

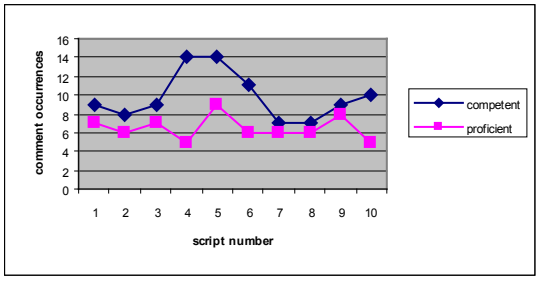


Figure 10.57. Comparing scripts to scale descriptor 1: layout when rating organisation

The attention to the second descriptor in the rating scale on links was different from the attention raters paid to the first scale descriptor. There were fewer comments, 151 (mean 4.1) than on layout features. Competent raters mentioned links 74 (mean 3.4) times and proficient raters evaluated links 77 (mean 5.1) times (see Table 10.50). Proficient raters' comments were more evenly distributed when they were comparing the scripts to the second scale descriptor (sd 1.83 and 2.12, respectively).

Table 10.50
Comparing Scripts to Scale Descriptor 2: Links When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A.4.b	Compares to scale descriptor 2: links	Competent		6	9	8	8	3	7	8	8	6	11	74 (3.4)	2.12
		Proficient		5	9	8	9	4	8	8	8	8	10	77 (5.1)	1.83
		Total		11	18	16	17	7	15	16	16	14	21	151 (4.1)	

The comment distribution patterns regarding links in the scripts were similar, as Figure 10.60 illustrates. Competent and proficient raters made numerous comments when they were evaluating the layout of script N5, but as far as linking text parts was concerned, they made the fewest comments. The most attention both groups of raters paid to script N10, competent raters made 11 and proficient raters 10 comments.

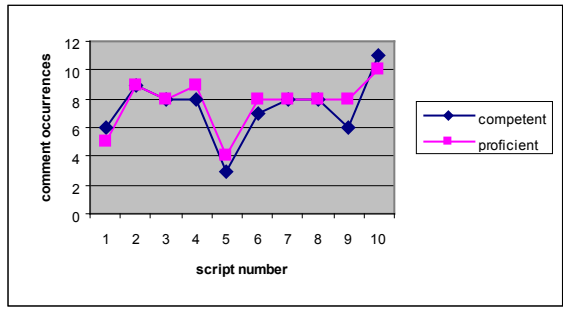


Figure 10.58. Comparing scripts to scale descriptor 2: links when rating organisation

In addition, raters read out the scale descriptors regularly, as Table 10.51 indicates. Competent raters read the scale descriptors more often 191 (mean 8.7) than proficient raters, who did so 94 (mean 6.3) times.

Table 10.51
Reading the Scale When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B4a	Reads scale	Competent		20	27	20	25	20	18	19	13	15	14	191 (8.7)	4.48
		Proficient		9	13	12	9	11	11	10	4	7	8	94 (6.3)	2.63
		Total		29	40	32	34	31	29	29	17	22	22	285 (7.7)	

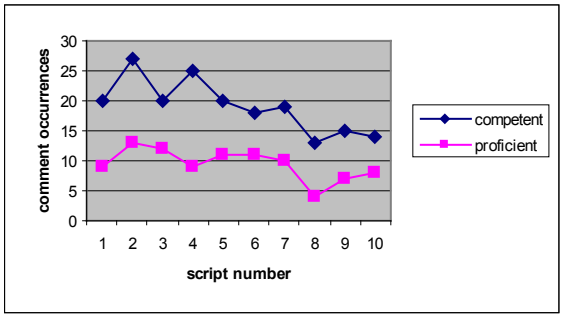


Figure 10.59. Reading the scale when rating organisation

Comment distribution curves, as it appears in Figure 10.59, are similar and both show a somewhat declining tendency. Competent raters' comments were more evenly distributed than those of proficient raters; however, they read the scale descriptors more often when they were evaluating scripts N2 and N4. Proficient raters read out the fewest descriptors when they were dealing with script N8.

Raters evaluated the aspect of organisation in their own words in 133 (mean 3.6) remarks, as Table 10.52 shows.

Table 10.52
Evaluation of Organisation in Own Words

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4c	Evaluates the aspects in own words	Competent		9	6	5	2	7	8	5	9	4	3	58 (2.6)	2.44
		Proficient		12	11	6	4	7	5	9	7	3	11	75 (5)	3.14
		Total		21	17	11	6	14	13	14	16	7	14	133 (3.6)	

Competent raters made considerably fewer such comments, 58 (mean 2.6) which were more evenly distributed across the ten scripts (sd 2.44) than comments by proficient raters (75 comments, mean 5).

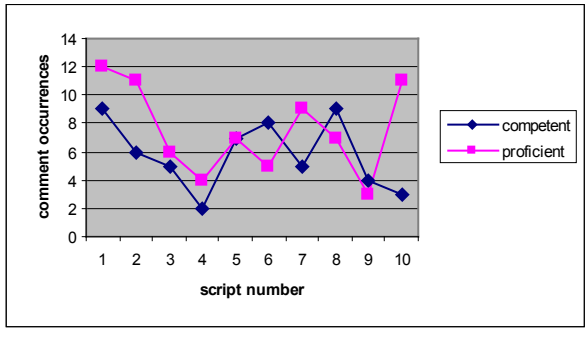


Figure 10.60. Evaluation of organisation in own words

The two distribution curves, as illustrated in Figure 10.60, are different regarding occurrences of comments on individual scripts. The biggest difference between competent and proficient raters' comments occurred on script N10: competent raters made 3 and proficient raters 11 comments. However, there were similarities between competent and proficient raters' attention to the scripts: for example, they evaluated script N1 in their own words considerably more than script N4.

Raters' additional criteria referred to individual details of different layout features: these comments were considered separately, as they constitute only an element of the criterion of layout. These own criteria were mostly remarks on paragraphing; out of the total 79 (mean 2.1) 50 which is 71% of all comments by competent and proficient raters (see Table 10.53 for details). The other additional criteria reflected on different elements of layout, such as appropriate wording for salutation and ending, or whether there was a date. There were some further remarks on sentence length, style and transitions which were not explicitly related to layout.

Table 10.53
Adding Own Criteria When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4d	Adds own criterion	Competent		3	4	7	4	8	1	5	7	11	6	56 (2.5)	2.84
		Proficient		0	1	2	2	2	2	4	5	3	2	23 (1.5)	1.42
		Total		3	5	9	6	10	3	9	12	14	8	79 (2.1)	

Competent raters added more of their own criteria, 56 (mean 2.5) than proficient raters, who added a criterion 23 (mean 1.5) times. Comments were more evenly distributed across the ten scripts for proficient raters (sd 1.42 and 2.84, respectively).

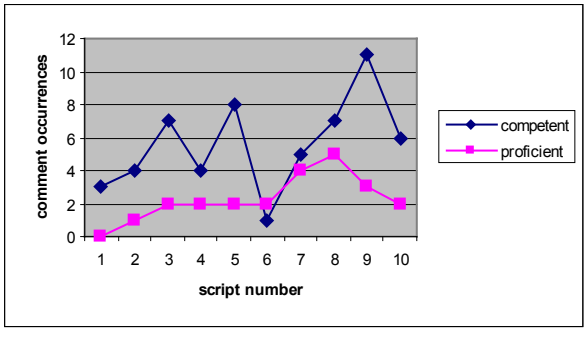


Figure 10.61. Adding own criteria when rating organisation

The biggest difference in comment occurrences on additional criteria by competent raters was between scripts N9 and N6, whereas the largest difference in occurrences appeared for scripts N8 and script N1 by proficient raters (see Figure 10.61 for details).

Raters almost always announced the score; there were 343 (mean 9.3) score nominations. Competent and proficient raters' attention was identical: competent raters made 204 (mean 9.3) and proficient raters 139 (mean 9.3) comments, respectively, as Table 10.54 shows.

Table 10.54
Score Nomination When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4e	Chooses score	Competent		18	19	21	22	21	20	21	21	21	20	204 (9.3)	1.17
		Proficient		14	13	14	13	14	14	14	10	16	17	139 (9.3)	1.85
		Total		32	32	35	35	35	34	35	31	37	37	343 (9.3)	

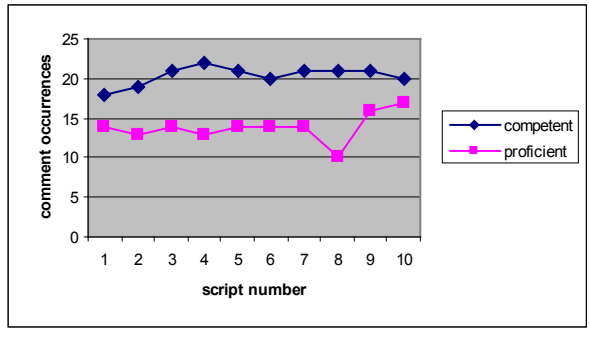


Figure 10.62. Score nomination when rating organisation

Distribution of the comments across the scripts indicates that some proficient raters failed to verbalise all scores: for example, when rating Script N8, and sometimes they announced the score more times, as with Script N9 and N10.

The additional strategies for rating the ten scripts comprise score justification, score revision, error identification, reference to lack of detail, focus change and score finalisation in all four rating criteria. Both competent and proficient raters made few additional comments when they were rating the aspect of organisation, as the detailed analysis below shows.

Raters rarely justified their scores: there were 46 (mean 1.2) occurrences. Competent raters more often added a reason, (31; mean 1.4) times and the comments were more unevenly distributed (sd 2.18), whereas proficient raters' 15 (mean 1) comments (see Table 10.55 for details) showed a more even distribution across the scripts (sd 1.65).

Table 10.55
Adding Reasons for Score Choice When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4f	Adds reason why that score	Competent		3	6	3	6	1	2	2	2	6	0	31 (1.4)	2.18
		Proficient		2	0	0	0	4	0	3	3	3	0	15 (1)	1.65
		Total		5	6	3	6	5	2	5	5	9	0	46 (1.2)	

There was a script, N10, for which none of the raters added a reason for score choice. The distribution of comments was very dissimilar in the two groups, as Figure 10.64 illustrates. Competent raters justified the scores with the most remarks for scripts N2, N4 and N9, whereas scores for scripts N2 and N4 were not justified by proficient raters at all. They paid the most attention to script N5 in this respect.

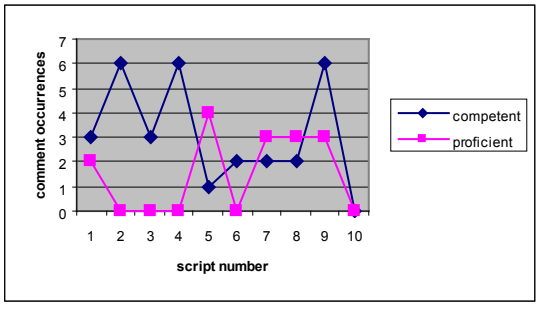


Figure 10.63. Adding reason for score choice when rating organisation

Raters very rarely revised their decisions, as Table 10.56 illustrates, there were three scripts, N2, N8 and N10 which were not mentioned and the decision for the other scripts was rarely revised, there were one or two comments. Although not the same scripts were mentioned, raters' attention was the same.

Table 10.56
Revision of Decision When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)
A4g	Revises decision	Competent		0	0	1	1	1	2	2	0	0	0	7 (.3)
		Proficient		1	0	1	0	1	0	0	0	1	0	4 (.3)
		Total		1	0	2	1	2	2	2	0	1	0	11 (.3)

Error identification was hardly used by raters when they were dealing with organisation, there were only three comments altogether, as Table 10.57 shows, competent raters identified an error in scripts N5 and N7, whereas a proficient rater found an error in script N8.

Table 10.57
Error Identification When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total
A4h	Identifies error	Competent		0	0	0	0	1	0	1	0	0	0	2
		Proficient		0	0	0	0	0	0	0	1	0	0	1
		Total		0	0	0	0	1	0	1	1	0	0	3

Raters sometimes remarked lack of detail when they were rating organisation: 35 (mean .9) comments. On a script, N10, none of the raters identified a lack of detail (see Table 10.58 for details). Competent raters noticed lack of detail fewer (20; mean .9) times than proficient raters (15; mean 1).

Table 10.58
Identification of Lack of Detail When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4i	Refers to lack of detail	Competent		1	2	1	4	3	0	3	2	4	0	20 (.9)	1.49
		Proficient		1	1	0	2	1	1	4	1	4	0	15 (1)	1.43
		Total		2	3	1	6	4	1	7	3	8	0	35 (.9)	

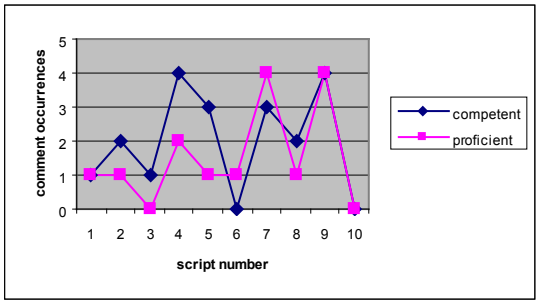


Figure 10.64. Identification of lack of detail when rating organisation

As far as comment occurrences across the ten scripts are concerned, there are similarities between the number of comments made by competent and proficient raters. Both groups commented on the most frequently on script N9 (4 comments each) and neither of them commented on script N10, as it appears in Figure 10.65.

Raters rarely changed focus when they were evaluating the organisation of the text, there were 26 (mean .7) remarks altogether, as Table 10.59 illustrates. Competent raters made much fewer such comments, only one or two for a script, than proficient raters, who made more (17; mean 1.1). There was a script, N5, which did not generate changing focus. Raters most often turned to the rating aspect of task achievement, out of the total of 26 comments, there were 11 (42%) in which competent and proficient raters switched to task achievement, in 7 (27%) they turned to the aspect of grammar, and the remaining comments were made on different other criteria, such as vocabulary, comprehension, relevance or overall impression.

Table 10.59
Changing Focus When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4j	Changes focus/ switches to different criterion	Competent		1	1	2	1	0	0	1	1	1	1	9 (.4)	.57
		Proficient		4	2	3	1	0	1	0	2	1	3	17 (1.1)	1.34
		Total		5	3	5	2	0	1	1	3	2	4	26 (.7)	

The distribution of comments across the ten scripts, as Figure 10.66 shows, was different for the two groups of raters: competent raters made the most comments on script N3 (two) and proficient raters on script N1 (four).

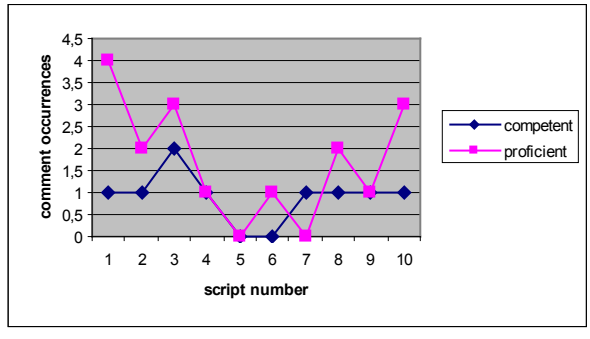


Figure 10.65. Changing focus when rating organisation

Proficient raters more often finalised scores: 46 (mean 3.1) times compared to competent raters (30; mean 1.4), as shown in Table 10.60.

Table 10.60
Score Finalisation When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
A4k	Finalises score	Competent		2	6	5	5	3	2	2	0	2	3	30 (1.4)	1.83
		Proficient		3	7	5	3	4	3	5	10	4	2	46 (3.1)	2.47
		Total		5	13	10	8	7	5	7	10	6	5	76 (2.1)	

The distribution of comment occurrences across the ten scripts was similar for competent and proficient raters except for script N8, whose score was finalised by competent raters in 10 comments, whereas none of the proficient raters finalised its score (see Figure 10.67 for details).

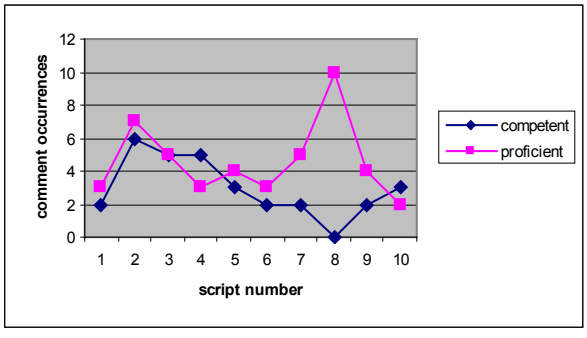


Figure 10.66. Score finalisation when rating organisation

Regarding reading the scripts, two subcategories were established in all four rating aspects: reading more words form scripts and reading out one-word examples.

Table 10.61
Reading More Words from Script When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B4b	Reads script: more words	Competent		0	0	6	3	4	4	2	4	1	2	26 (1.2)	1.96
		Proficient		4	7	2	3	2	3	2	5	5	9	42 (2.8)	2.35
		Total		4	7	8	6	6	7	4	9	6	11	68 (1.8)	

As far as reading more words from the scripts is concerned, raters read the scripts extensively 68 (mean 1.8) times, as demonstrated in Table 10.61. Competent raters turned to this strategy less frequently; they made 26 (mean 1.2) comments, while proficient raters made 42 (mean 2.8) comments.

The comment distribution curves, as illustrated in Figure 10.68 show different tendencies, for example, competent raters did not read text when they were evaluating scripts N1 and N2 for organisation, while proficient raters read from texts, especially from N2, which received the most attention. The other considerable difference was observed in connection with script N10, which was commented nine times by proficient raters and twice by competent raters.

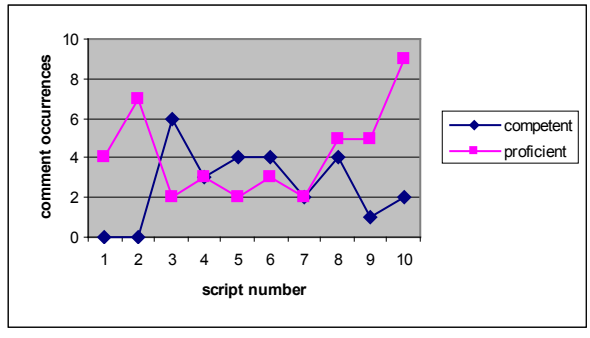


Figure 10.67. Reading more words from script when rating organisation

Findings of reading one-word examples from the scripts show the opposite tendency, as indicated in Table 10.62. Competent raters read out individual words more often (49; mean 2.2) than proficient raters (9; mean .6).

Table 10.62
Reading One-word Example When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)	sd
B4d	Reads example: one word	Competent		5	8	2	2	3	6	2	4	6	11	49 (2.2)	2.96
		Proficient		2	4	0	0	0	0	0	2	0	1	9 (.6)	1.37
		Total		7	12	2	2	3	6	2	6	6	12	58 (1.6)	

The distribution of comments was somewhat similar in tendency, as Figure 10.69 illustrates; however, competent raters read not only considerably more from script N10 than proficient raters did, but they read much more than from any other script.

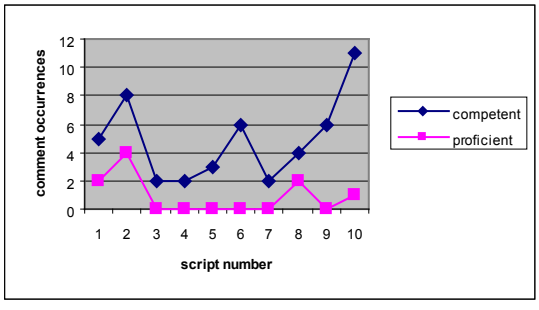


Figure 10.68. Reading one-word examples when rating organisation

Raters almost never summarised content, there were 16 (mean .4) comments altogether, as Table 10.63 illustrates, out of which competent raters summarised content of scripts N2, N4 and N9 once. Proficient raters, on the other hand, mentioned script N2 five times and they made one or two comments on six other scripts.

Table 10.63
Summarising Script When Rating Organisation

Code	Comment type	Rater	Script	1	2	3	4	5	6	7	8	9	10	Total (mean)
B4c	Summarises script	Competent		0	1	0	1	0	0	0	0	1	0	3(.1)
		Proficient		0	5	1	0	0	2	1	1	2	1	13(.9)
		Total		0	6	1	1	0	2	1	1	3	1	16(.4)

Regarding raters' reference to the rubric, three proficient raters referred to rubric when rating organisation.

Raters' focus on organisation was similar in some cases and different in others: they attended to certain features more often than to others and there were differences in their attention to the ten scripts:

e. Competent and proficient raters attended to the rating scale descriptors closely when they were evaluating the rating criterion of organisation, but there were differences between the two groups. Proficient raters seemed to pay more attention to the logical structure of texts; they made more comments on links in the texts. As for reading the scale descriptors, competent raters turned to this strategy more often than proficient raters did. In addition, they read more from the scale at the beginning of the rating process than at the end.

f. Raters sometimes rated organisation in their own words, especially proficient raters made much more comments in this respect than competent raters.
g. Both competent and proficient raters revised their decision strategy on few occasions, which can lead to the tentative conclusion that they felt comfortable with the evaluation of the criterion and did not need to change their score. Error identification was not characteristic when raters were evaluating the rating criterion of organisation. As far as additional criteria are concerned, some raters narrowed down the rating criterion of layout to one of its components, which was a specific feature for a letter and they focused on it. Similarly, they mainly switched to of task achievement when they were rating organisation and remarked the way the script resembled an informal letter.
h. Findings show that raters frequently attended to the scale when rating the criterion of organisation. However, they did not use many of the reading strategies and almost never read the task rubric.

10.6 Conclusion

The present chapter focused on raters' interpretation of rating criteria and I compared competent and proficient raters' focus on task achievement, vocabulary, grammar and organisation. Investigation into similarities and differences between the two groups of raters was possible, because their verbalised rating processes were categorised based on the same set of criteria. Chapter Nine centred on the examination of raters' foci on the four rating criteria and findings showed that the most attention was paid to rating strategies and competent raters focused more on rating strategies than proficient raters. Reading behaviour comments were also investigated, which were fewer than rating strategies and competent raters turned to reading strategies more often than competent raters did. Attention to the four rating criteria was different: grammar attracted the most attention of both groups of raters; however, their attention to task achievement, vocabulary and organisation showed different patterns. I investigated these emerging rating patterns further in this chapter to answer the third research question on the way competent and proficient raters interpreted the four rating criteria.

In order to answer the third research question, four specific questions were formulated about raters' attention to (a) the scale descriptors, (b) the rating criteria, (c) additional criteria and (d) the scale, the scripts and the task rubric.

On the four rating criteria, some patterns emerged that raters followed when they were evaluating the ten scripts.

a. Competent and proficient raters attended to both scale descriptors when

they were evaluating the ten scripts, but their attention was different on the individual aspects. When rating task achievement both groups of raters were more engaged in rating content points than the achievement of communicative goal. Competent raters attended to content points more and they read out more from the scale. The pattern for rating vocabulary was different, as competent raters attended less to the scale descriptors than proficient raters did. However, the number of comments they made when comparing scripts to the scale descriptor was similar for both rater groups. As far as reading is concerned, competent raters read the scale more, as in case of task achievement. Another pattern emerged when raters were dealing with grammar: competent raters attended more to accuracy and less to structures than proficient raters did. Reading the scale comments were distributed similarly to reading when dealing with task achievement and vocabulary: competent raters read the scale more often than proficient raters did. As for organisation, regarding the first descriptor on layout features, both groups of raters considered it similarly, however, the second descriptor on evaluating links attracted proficient raters' attention more. Competent raters turned to reading the scale descriptors more frequently.

b. Evaluation of the rating aspect in raters' own words was carried out considerably more often by proficient raters on each of the four aspects, the difference was especially large when they were rating the aspect of grammar and organisation.
c. Although there was some observable difference between the occurrences of additional comments by the two groups of raters on the four rating criteria: few comments occurred and they were similar. The additional criteria that competent and proficient raters turned to mostly included some of the details of the aspect, which was most apparent when organisation was evaluated. Regarding the comment occurrences, proficient raters made more comments in this subcategory, except for organisation.
d. Competent and proficient raters paid considerable attention to the four rating criteria, as discussed above; however, they turned to the scale descriptors differently, and sometimes they paraphrased the evaluation. As far as attention to scripts is concerned, competent raters read considerably more from the text, especially when rating grammar, and sometimes cited one-word examples as well. Extended text reading behaviour comments occurred when rating grammar and task achievement and the least when rating organisation. Competent and proficient raters' patterns were different across the four rating aspects: competent raters read considerably more when they were evaluating task achievement, whereas proficient raters read more when they were dealing with grammar and vocabulary. Both groups of raters read the least when they were rating organisation. One-word examples were more frequent for vocabulary and grammar and

rare for task achievement and organisation. They paid the least attention to the rubric, there were no remarks on reference to the rubric when rating grammar and there were only five comments when raters were dealing with vocabulary. Some raters mentioned the rubric when rating organisation and paid the most attention to it when they evaluated task achievement.

There were differences not only between competent and proficient raters' focus on each of the four rating criteria of task achievement, vocabulary, grammar and organisation, but also in the attention they paid to individual scripts when rating them according to each criterion. These differences are examined in what follows by looking into the way the weakest script, N2 and the top script, N6 were rated by competent and proficient raters.

Chapter 11

Raters' Script Interpretation on the Weakest and Top Script

Introduction

In Chapter Eight I discussed rating sequences and emerging rating patterns that raters develop, whereas in Chapters Nine and Ten I investigated what competent and proficient raters focused on. These features contribute to a deeper understanding of written performance assessment characteristics comprising raters, rating scale, rating processes, students' performance, task or instrument and candidate characteristics (McNamara, 1996). The present chapter centres on the relationship between raters' rating behaviour and students' performance and tries to answer the following research question:

4. How do competent and proficient raters interpret the scripts?

Rating is influenced by the raters' understanding of the texts and the rating criteria. When raters are reading the scripts, they are trying to comprehend what writers intended to say. During this process raters interpret the text somehow, they create a certain image of the text which they rate afterwards (Wolfe, 1997).

In order to find the answer to the fourth research question I will look into rating processes of the weakest and top scripts to understand how raters interpret this particular script. According to the benchmarks, Script N2 was the weakest and script N6 represented the top performance among the ten scripts. Photocopies of the original scripts can be found in Appendix 7.4. First, in this chapter, I discuss script N2, the weakest out of the ten scripts, analyse the rating processes and compare scores awarded by raters as well as the benchmarks; then, I detail the assessment of script N6 along using the same procedure to elaborate on raters' script and scale interpretation. Starting with the pre-scoring stage, raters' rating sequences are detailed. The four rating criteria are dealt with separately in the order as they appear in the rating scale. Finally, a comparison of the awarded scores and raters' verbalised evaluation of the script is made to shed light on raters' decision-making processes.

11.1 Rating Script N2: Benchmarks and Total Scores

The scores express raters' decisions in numbers, and they are scores assigned according to descriptors in a rating scale. An analytic rating scale, as described in Chapter Four, comprises descriptors included in bands to distinguish between performance levels. Descriptors are devised by test designers on the basis of their knowledge about the construct in question and test design principles. Raters go through a rater training before the rating takes place and carry out the rating task during which they choose scores from the scale. Each element in performance assessment has its characteristics and plays a role in the raters' decision making process. The relationship between the scores awarded, rating processes and raters' thinking is examined in detail to find out more about the way raters make their decisions.

According to the benchmarks for the ten scripts, as presented in Chapter Seven in Table 7.2, script N2 was ranked last with a total of 9 points. Competent and proficient raters' rankings were somewhat different. The total scores and competent raters' rankings on the ten scripts are in Appendix 11.1 and proficient raters' total scores and their rankings are in Appendix 11.2. Six (27%) competent raters ranked script N2 last and twelve (55%) decided that script N4 was the weakest. However, eight (53%) proficient raters chose script N2 as the weakest and six (40%) chose script N4 as the weakest.

Looking at the scores awarded on each of the four criteria, where the top score was 6 points, adding up a total of 24 points the maximum total score (see Appendix 7.2 for the rating scale), we can see that all total mean scores awarded for script N2 were higher in each category than the benchmark. The competent raters' total score was higher (13.14 points; sd 4.56) than that of proficient raters (11.53 points; sd 3.60), which means that proficient raters' scores were closer to the benchmark (see Table 11.1 for details). Organisation was considered to be the weakest of the four rating criteria by the researcher, who awarded a 1 as a benchmark. However, both competent and proficient raters considered task achievement the weakest out of the four criteria: their mean scores were 2.77 (sd 1.31) and 2.33 (sd 1.35), respectively. Similarly to the benchmarks on vocabulary and grammar, both groups of raters' mean scores were similar and the highest on these two criteria. Proficient raters awarded consistently lower scores to the script than competent raters and competent raters' standard deviation figures show bigger differences between the scores.

Table 11.1
Benchmarks and Mean Scores on Four Rating Criteria of Script N2

	Task achievement	Vocabulary	Grammar	Organisation	Total
Benchmarks	2	3	3	1	9
All raters' mean (sd)	2.58 (1.31)	3.71 (1.33)	3.50 (1.22)	2.61 (1.17)	12.39 (4.20)
Competent raters' mean (sd)	2.77 (1.31)	3.95 (1.46)	3.95 (1.40)	2.82 (1.18)	13.14 (4.56)
Proficient raters' mean (sd)	2.33 (1.35)	3.40 (1.12)	3.40 (.99)	2.40 (1.12)	11.53 (3.60)

To sum up, script N2 was one of the very weak performances according to all 37 raters; however, competent raters awarded somewhat higher scores than proficient raters. All raters agreed that task achievement was the weakest of all criteria, which can be partly justified by overall comments on relevance of the text, as discussed below, in the pre-scoring stage of rating. In order to find out more about the reasons for these differences, I examine the rating processes in the next sections. The analysis follows the sequence most raters went through during rating: first the pre-scoring stage, and then the four rating criteria one by one in the order they were considered.

11.1.1 Comments on Script N2 in the Pre-Scoring Stage

As mentioned earlier in Chapter Eight, some raters started the rating process with making overall comments with different foci. This first, initial stage in rating is referred to in earlier studies as pre-scoring stage (Milanovic et al., 1996; Lumley, 2000; 2002). When rating script N2, seven competent raters (32%) and ten proficient raters (67%) started with overall comments. They focused on surface features, such as layout and legibility, as examples of a competent rater's protocol in Excerpt 11.1 show.

Excerpt 11.1: R1's pre-scoring stage comments on script N2

R1

TU	Rater talk
1	this composition is short, I think,
2	and it is difficult to read

Competent and proficient raters both mentioned surface features; they all thought that the composition was short and difficult to read due to its messy layout.

The overall impression with content focus however, represent two extremes in both groups. There were two raters, one from each group, who expressed their appreciation of the content and made positive comments, such as the example of a proficient rater shows in Excerpt 11.2. However, she awarded moderate scores (2, 4, 4 and 4) when later she was looking at each criterion.

Excerpt 11.2: R8's positive comment on content in the pre-scoring stage of script N2

R8

TU	Rater talk
1	starts pretty well

The other extreme was the raters' negative comments: they remarked the irrelevance of the content saying that they did not understand the text. There were three raters, one competent and two proficient ones, who gave a 0 for task achievement and they said that the student misunderstood the task. For example, as a competent rater's example in Excerpt 11.3 demonstrates, she stated at the beginning that the text was irrelevant.

Excerpt 11.3: R5's remark on text irrelevance in the pre-scoring stage of script N2

R5

TU	Rater talk
1	I think he misunderstood the task totally

Then she went on and summarised the content and made two comments on relevance before awarding a 0 and finishing rating at this point, as the scale did not have descriptors for other criteria in this band.

Another competent rater, R4, first compared the script to the previously rated ones, then meditated on student's proficiency and concluded on the writer of the letter, as it appears in Excerpt 11.4.

Excerpt 11.4: R4's remarks on text irrelevance in the pre-scoring stage of script N2

R4

TU	Rater talk
2	there is a kind of spanning of thoughts from one another at the beginning, I just felt at the end that he added something not totally relevant there
3	as far as the letter is concerned, my first impression was that the student had been to the US, or watched a lot of movies, his style reminds me of American English, he writes in a very casual and colloquial style

The other comments regarding script content were similar for competent and proficient raters: apart from comments on student's misunderstanding of the task, raters referred to the style, saying that it was too informal, or said that the text was strange or it was difficult to make sense of.

Apparently, the first overall impression did not seem to affect the raters, as, for example, the two raters who made positive comments were not influenced in their rating, as they awarded low scores on the rating criteria.

11.1.2 Rating Task Achievement of Script N2

Task achievement was rated first of the four rating criteria by most raters and as the scores show, competent raters chose a score from a wider range than proficient raters (see Table 11.2 for details). Competent raters chose a score from the whole range, from 0 to 6. However, they mostly chose a 2: eight raters (36%) or a 3: seven raters (32%). On the other hand, proficient raters mostly chose a 2: six raters (40%) or a 4 (47%) for task achievement and more of them gave 5 or 6 points.

Table 11.2
Script N2 Task Achievement Score Frequencies (benchmark column highlighted)

Task achievement score	0	1	2	3	4	5	6	Total
Frequencies (percentages)								
Competent	1 (5%)	1 (5%)	8 (36%)	7 (32%)	3 (13%)	1 (5%)	1 (5%)	22
Proficient	2 (13%)	1 (7%)	6 (40%)	2 (13%)	4 (27%)	0	0	15
Total	3 (8%)	2 (5%)	14 (38%)	9 (24%)	7 (19%)	1 (3%)	1 (3%)	37

Two proficient raters (13%) found the text irrelevant and awarded a 0 and two of them (13%) gave a 3 and one (7%) rater decided on score 1 for the criterion.

To sum up, we can see that both groups of raters chose a score from a wide range of scores and there were two competent raters who rated the script much higher than the others and there were three raters (one competent and two proficient) found the text irrelevant and gave the score of 0 for task achievement.

As far as rating processes are concerned, competent raters made 75 rating behaviour related comments (mean 3.4), whereas proficient raters 41 (mean 2.7), as Table 11.3 demonstrates. Raters' reading strategy use was very different: competent raters read 83 (mean 3.8) times and proficient raters 19 (mean 1.3) times. Own focus comment occurrences show that competent raters turned to their own criteria less frequently (40 comments; mean 1.8), which was 25% of all 198 (mean 5.3) their remarks on script N2. Proficient raters made a total of 22 (mean 1.5) own focus comments when they were rating task achievement which was 37% of a total of 82 (mean 2.2). These figures show that competent raters paid more attention to evaluating task achievement than proficient raters. The biggest difference was in reading strategies between the two groups of raters (see Figure 11.1 for details).

Table 11.3
Rating Foci Comments When Rating Task Achievement of Script N2

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	75 (3.4)	83 (3.8)	40 (1.8)	198 (5.3)
Proficient	41 (2.7)	19 (1.3)	22 (1.5)	82 (2.2)
Total	116 (3.1)	102 (2.8)	62 (1.7)	280 (7.6)

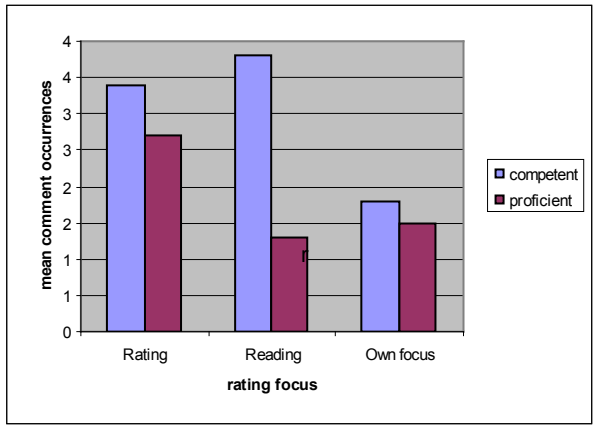


Figure 11.1. Rating foci comments when rating task achievement of script N2

Neither proficient nor competent raters used the "compares script to the first scale descriptor" strategy for rating achievement of communicative goal criterion, as illustrated in Table 11.4.

Table 11.4
Comments When Rating Task Achievement of Script N2

	Codes											
Raters	A1a	A1b	A1c	A1d	A1e	A1f	A1g	A1h	A1i	A1j	A1k	Total (mean)
Competent	0	18	3	0	20	3	3	0	21	3	4	75 (3.4)
Proficient	0	8	1	1	13	1	0	0	10	0	7	41 (2.7)
Total	0	26	4	1	33	4	3	0	31	3	11	116 (3.1)

The achievement of communicative goal seemed to be less important for raters than the number of content points covered. Almost all competent raters checked the number of content points: they did so 18 times and compared the script to the second scale descriptor, whereas eight proficient raters made at least one explicit comment in this category.

Raters indicated when they were choosing a score, 20 competent and 13 proficient raters announced the score they chose. Apart from these, the most frequently observed comment in rating task achievement was reference to lack of detail: 13 competent raters made a total of 21 comments and six proficient raters made a total of ten comments saying that some topics or content points were not mentioned in the texts.

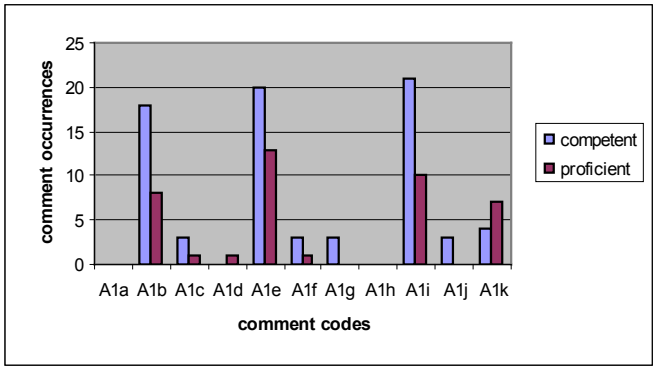


Figure 11.2. Comment occurrences when rating task achievement of script N2

The other strategies occurred once in the proficient group (see Figure 11.2), for example, only one proficient rater paraphrased the criterion of task achievement, another added an own criterion and three competent raters paraphrased the criterion.

The same number, three competent raters justified their decision, revised their decision or changed focus when rating the script. In addition, there were four and seven occurrences of score finalisation, respectively.

Looking at the total number of reading behaviours, competent raters read considerably more (83 comments; mean 3.8), while proficient raters read on 19 (mean 1.2) occasions (see Table 11.5 for details).

Table 11.5
Reading Related Comments When Rating Task Achievement of Script N2

	Codes					
Raters	B1a	B1b	B1c	B1d	B1e	Total (mean)
Competent	21	25	14	4	19	83 (3.8)
Proficient	9	4	5	0	1	19 (1.3)
Total	30	29	19	4	20	102 (2.8)

The distribution of competent and proficient raters' reading strategies demonstrates that competent raters relied on reading much more in their evaluation than proficient raters, as it appears in Figure 11.3.

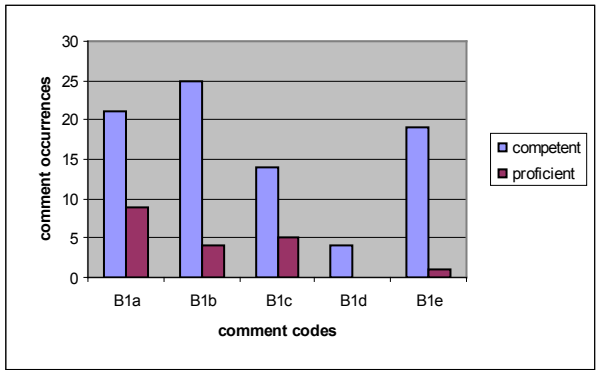


Figure 11.3. Reading related comments when rating task achievement of script N2

The first reading behaviour was reading the rating scale descriptors and competent raters turned to the scale more often than proficient ones. Comment

occurrences in reading the text category show bigger differences as far as the number of comments is concerned: four proficient raters read parts of the text, whereas eight competent raters read parts of the text with a total of 25 occurrences, as some of them read text parts not only once; RR5 read the text the most often (ten times). Nine competent raters summarised script content 14 times and four proficient raters did so five times altogether. None of the proficient raters cited one-word examples when rating task achievement, but two competent raters quoted four words in this category.

Competent raters often referred to the rubric, while there was one out of the 15 proficient raters who read the rubric, as Excerpt 11.5 illustrates.

Excerpt 11.5: RR3's reading the task rubric when rating task achievement of script N2

RR3

TU	Rater talk
11	the **invite yout friend for next holiday** is only
12	mentioned as *Can you come?*

On the other hand, five competent raters referred to the rubric 19 times, so their rating pattern for task achievement was different from the others' rating patterns, as they compared parts of text to the rubric while evaluating task achievement. There is an extract from one of these raters' protocol in Excerpt 11.6.

Excerpt 11.6: R10's reading the rubric when evaluating task achievement of script N2

R10

TU	Rater talk
1	now, let's have a look ... **nice time** done
2	*service*
3	*stomach-ache,*
4	present, **the journey home** ... aha
5	yes, got it
6	*plain* ... ok
7	**present**

These raters followed a kind of an inventory pattern; they checked which content points were covered and compared the rubric to the text.

Regarding own focus, the main concern for both competent and proficient raters seemed to be content relevance, which in this case meant irrelevance,

as all raters referred to irrelevance of the text: eight (36%) competent raters mentioned content relevance eleven times and five (33%) proficient raters pointed at text relevance, as the example in Excerpt 11.6 illustrates.

Excerpt 11.7: R1's comment on text relevance when rating task achievement of script N2

R1

TU	Rater talk
10	the writer didn't write about what he had to, but he wrote some jokes

Although this proficient rater mentioned irrelevance in the pre-scoring stage of her rating process, such comments did not occur in other sections of raters' protocols. It implies that raters referred to content relevance in the pre-scoring stage or associated it with task achievement.

In addition to the comments on relevance, raters' concern about the content was expressed in a high number of reflections of raters' feelings. Competent raters made eleven such comments, which is 28% of the total of 40, and some proficient raters expressed their own feelings eight times, which is 36% of the total of 22. An example from a competent rater's protocols in Excerpt 11.8 illustrates reference to feelings.

Excerpt 11.8: RR14's own feelings related to task achievement of script N2

RR14

TU	Rater talk
4	[he] is rather meditating, and ... communicating different thoughts instead of concentrating on things prescribed in the instructions

Proficient raters expressed their feelings similarly, as the example in Excerpt 11.9 shows.

Excerpt 11.9: RR11's own feelings related to task achievement of script N2

RR11

TU	Rater talk
5	and I think there are some illogical statements in the letter

As far as the awarded scores are concerned, those raters' processes are mainly examined, whose score agrees with the benchmark score of two points: there

were eight (36%) competent and six (40%) proficient raters (see Table 11.2 for details) who awarded two points for task achievement. Score distribution in Figure 11.4 shows that competent raters chose a score from a wider range than proficient raters, whose score was somewhat higher than the benchmark.

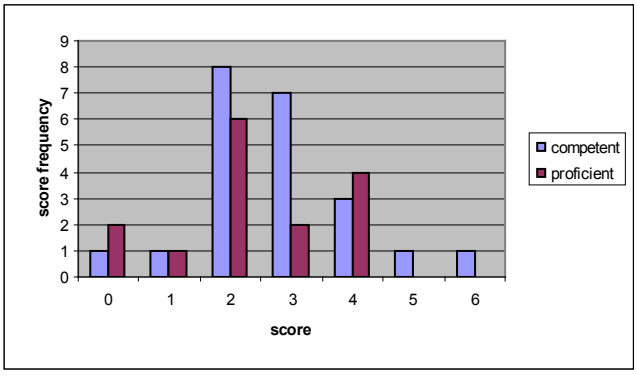


Figure 11.4. Scores on task achievement of script N2

There were two raters though, who seemed to contradict the rating scale contents: one read out a scale descriptor for the band of 4 points and a negation of another one, the top band and still awarded 2 points, as Excerpt 11.10 demonstrates.

Excerpt 11.10: R16's rating task achievement of script N2

R16

TU	Rater talk
3	so he **almost covered all the content points**,
4	but he **could not achieve the communicative goal**
5	I think, it's not really comprehensible,
6	he does not invite his friend for next holiday
7	but he just drops a few lines about a party next week and how could a friend come from England for next Friday.
8	And nothing is mentioned about his plans about future and about the places they would visit
9	and the part about this water all around is messy and it cannot be comprehended
10	I think, it wants o be humorous but I do not really think so,
11	I gave two for

Another rater stated twice that the student totally misunderstood the task, which is the descriptor for the bottom band of 0 and, still, awarded 2 points (see Excerpt 11.11 for the rater's words).

Excerpt 11.11: R2's rating task achievement of script N2

R2

TU	Rater talk
5	It seems that he did not understand the task at all, as ... mostly writes about himself and a journey
6	So 2 points for task achievement
7	But, maybe I have to correct myself as far as task achievement is concerned, because I think he interpreted the task as if he moved somewhere, perhaps home
8	I don't know exactly
9	But still, he wrote about things required
10	All the same, let it be 2

Looking at the way competent raters arrived at 2 points as a score for task achievement, we can observe that the number of content points covered played a crucial role in raters' decision, as most competent raters who awarded 2 points for task achievement referred to two or three content points that were covered in the texts.

In contrast, it seems as if R9, a competent rater had a completely different idea of the script, as he awarded the top score, 6 points for task achievement and evaluated it reading out both descriptors in the rating scale for top performance. Apart from reading out the two descriptors from the scale, as is illustrated in Excerpt 11.12, there is no other reference to task achievement in his evaluation.

Excerpt 11.12: R9's rating task achievement of script N2

R9

TU	Rater talk
1	Now let's see the next one which is script number 2
2	And after the first reading,
3	it seems that the content points here also... all four... **all five content points were covered**
4	**and the communicative goal was also achieved**
5	in this case, so for task achievement
6	it is 6.

Similarly, R18, who gave 5 points for task achievement verbalised her evaluation and said that all content points were covered and she did not refer to content, just confirmed that something was written, as it appears in Excerpt 11.13.

Excerpt 11.13: R18's rating task achievement of script N2

R18

TU	Rater talk
5	because it is true that **all 5 content points are covered**
6	but there are only two sentences one each for the last two about the invitation and programmes
7	but he inserted ... a topic
8	That's why it is only 5

These two raters obviously did not pay attention to the content of the script; they treated the text as if it corresponded to the rubric requirements, so they considered the task accomplished.

Three raters thought the text was irrelevant in the pre-scoring stage, and then when they were comparing the script to the rating criteria they decided to award a 0, as Excerpt 11.14 demonstrates. These raters concentrated on the content of the text and attempted to make sense of it but as they could not, they scored it 0.

Excerpt 11.14: R5's rating task achievement of script N2

R5

TU	Rater talk
1	I think he totally misunderstood the task
2	He is not writing about a holiday and not that he spent some time with his friend, but is talking about moving. Brings in somehow ... talk about a kind of travelling. Travelled on a plane on the first class. And that he bought some kind of pyjamas.
3	But what has it got to do with a party invitation for next Friday?
4	In addition, with life that is so different here from that of back home.
5	He completely ... he is writing completely about something else not what was needed.
6	Zero, in all aspects zero.

To sum up, observation in connection with rating task achievement of script N2 is that raters paid more attention to the fulfilment of content points than to the achievement of the communicative goal. It is especially so with competent raters, who awarded scores without an apparent attention to the content of the letter: they evaluated the scripts on the basis of comparison of the text with the

content points regardless of what it was about. However, frequent reference to students' misunderstanding of the task contradicts this observation and raters seemingly tried to do their best to interpret the text in a way so as to match it to the rating criteria.

11.1.3 Rating Vocabulary of Script N2

After rating the task achievement, raters dealt with the second criterion: vocabulary. Looking at score distribution, as shown in Table 11.6, in which the benchmark column is highlighted, we can see that more competent raters gave a score of 4 or 5 rather than 3 points, which was the benchmark. Most proficient raters' scores were distributed within the range of 2 and 4 points. Fewer proficient raters' scores were very different from the benchmark than those of competent raters. Altogether nine raters (24%) awarded the same score as the benchmark: four competent raters' (18%) and five proficient (33%) ones' score was identical with the benchmark. As the mean scores on vocabulary in Table 11.1 show, competent raters' vocabulary mean score was higher than that of proficient raters', the mean scores were 3.95 (sd 1.46) and 3.40 (sd 1.12), respectively.

Table 11.6
Script N2 Vocabulary Score Frequencies (benchmark column highlighted)

Vocabulary score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Frequencies (percentages)								
Competent	1 (5%)	0	2 (9%)	4 (18%)	7 (32%)	5 (23%)	3 (14%)	22
Proficient	0	0	4 (27%)	5 (33%)	4 (27%)	1 (7%)	1 (7%)	15
Total	1 (3%)	0	6 (16%)	9 (24%)	11 (30%)	6 (16%)	4 (11%)	37

Now the details of rating are elaborated, starting with rating processes, and then a comparison of awarded scores to verbalised evaluations is carried out.

Rating vocabulary was characterised by a different number of comments with a rating and reading focus: raters made somewhat more reading (114; mean 3.1) than rating (113; mean 3) related comments, as Table 11.7 shows. In addition, they referred to their own criteria 44 (mean 1.2) times, competent raters less frequently than proficient ones (23; mean 1.1 and 21; mean 1.1, respectively). Looking at the number of rating and reading related comments, we can see that competent raters made fewer remarks on this rating criterion.

Table 11.7
Focus When Rating Vocabulary of Script N2

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	54 (2.5)	64 (3)	23 (1.1)	141 (6.4)
Proficient	59 (3.9)	50 (3.3)	21 (1.4)	130 (8.7)
Total	113 (3)	114 (3.1)	44 (1.2)	271 (7.3)

Considering the comment means, competent raters turned to rating strategies more (54; mean 2.5) than to reading (64; mean 3). Proficient raters' focus was the opposite: they made 59 (mean 3.9) rating focus comments and 50 (mean 3.3) reading related ones.

The distribution of competent and proficient raters' comment means on the three rating foci is illustrated in Figure 11.5, which shows that competent raters focused the most on reading and proficient raters on rating strategies.

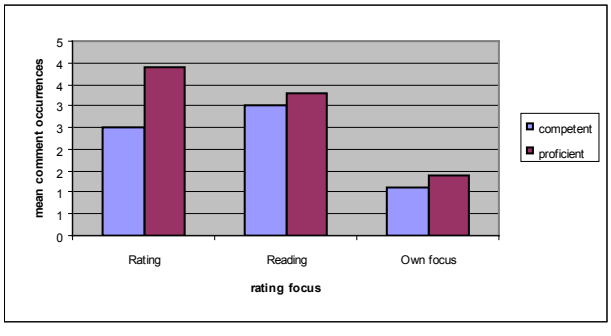


Figure 11.5. Rating foci when evaluating vocabulary of script N2

Proficient raters made more of their own focus comments than competent raters.

Regarding raters' rating focus, as Table 11.8 shows, there were 54 (mean 2.5) rating remarks by competent raters and 59 (mean 4) by proficient raters and their attention to the rating criteria was differently distributed.

Table 11.8
Rating Focus for Vocabulary of Script N2

	Codes											
Raters	A2a	A2b	A2c	A2d	A2e	A2f	A2g	A2h	A2i	A2j	A2k	Total (mean)
Competent	4	4	9	3	21	3	1	4	0	2	3	54 (2.5)
Proficient	3	6	8	9	13	4	2	3	0	7	4	59 (4)
Total	7	10	17	12	34	7	3	7	0	9	7	113 (3)

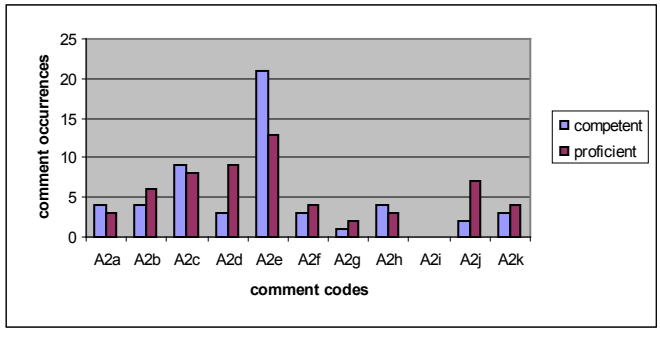


Figure 11.6. Rating comments for vocabulary of script N2

Four (18%) competent raters referred to vocabulary range and four (18%) to appropriacy when they compared the script to scale descriptors. Only one competent rater, R17, mentioned both scale descriptors. As far as proficient raters are concerned, three (20%) noted range and three (20%) made six comments on appropriacy and one talked about both scale descriptors.

Vocabulary criterion was evaluated in own words by nine (40%) competent raters and eight (53%) proficient raters, as an example of a competent rater (TU13) and a proficient one (TU11) in Excerpt 11.15 shows.

Excerpt 11.15: R15's and RR1's rating vocabulary of script N2 in own words

R15

TU	Rater talk
12	**basic words and expressions**
13	and sometimes irrelevant

RR1

TU	Rater talk
11	Well vocabulary that he uses is interesting, and that's why as there are interesting things in it, these similes and he uses some words, such as...
12	*gorgeous*

Raters added their own criteria when rating vocabulary, three (14%) competent raters referred to spelling mistakes and seven (47%) proficient made nine similar comments on spelling words. Some raters remarked word order, but not when evaluating vocabulary, as RR15 did. Apart from these criteria, a rater, R14 made a remark on style (see Excerpt 11.16).

Excerpt 11.16: RR15's and R14's additional criteria for rating vocabulary of script N2

RR15

TU	Rater talk
4	He is using very **basic words and expressions**
5	and even in them he has a lot of mistakes

TU	Rater talk
6	he mixes the order of words in sentences

R14

TU	Rater talk
17	Mmm ... pretty good, let's say, pretty good. The letter is casually composed, belongs to informal letters

As far as choosing a score is concerned, those raters failed to announce the score they chose, who did not voice their choice for task achievement. The only difference in frequencies was that there was one competent rater announced the score twice. The next rating behaviour type was score justification, three (14%) competent and three (20%) proficient raters justified their scores, one proficient rater did so twice. Raters rarely revised their decisions, there was one such a comment by a competent and two by proficient raters.

The two comment types related to errors were error identification which was classified within rating decisions, and the other, error correction which was among own focus comments. Four (18%) competent raters identified errors and three (20%) proficient raters said that a mistake or mistakes were made in the texts. When rating vocabulary, none of the raters referred to a lack of detail, it

seems that they did not identify this behaviour type with vocabulary evaluation. Changing focus occurred in two (9%) competent raters' and in four (27%) proficient raters' protocols; the latter made a total of 7 comments in this respect. The two examples in Excerpt 11.17, one from a competent and a proficient rater each to show typical comments, as R16 referred to problems with the content when evaluating vocabulary and content evaluation belonged to the criterion of task achievement. The proficient rater, R14, remarked on sentence structures when rating vocabulary. Score finalisation occurred in two (9%) competent and in four (27%) proficient raters' protocols.

Excerpt 11.17: R16's and R14's changing focus when evaluating vocabulary of script N2

R16

TU	Rater talk
17	there are inaccuracies in the text for example,
18	how could a friend come from England for next Friday - he should ask about next holiday

R14

TU	Rater talk
20	Sentences are compiled relatively well

As for reading behaviour comments, competent raters read 64 (mean 3) times, and proficient raters 50 (mean 3.3) times, as Table 11.7 above shows. Reading behaviour related comments comprised reading the scale, more words or one-word examples from the script, text summary and reading the rubric. Competent and proficient raters were reading the scale and the texts differently, as Table 11.9 demonstrates.

Table 11.9
Reading Comments When Rating Vocabulary of Script N2

	Codes					
Rater	B2a	B2b	B2c	B2d	B2e	Total (mean)
Competent	16	21	2	24	1	64 (3)
Proficient	10	23	0	17	0	50 (3.3)
Total	26	44	2	41	1	114 (3)

The first was reading the scale descriptors: twelve (55%) competent raters read the scale 16 times and seven (47%) proficient raters did so ten times.

Text parts were read out by ten (45%) competent raters in 21 cases, and eight (53%) proficient raters read a total of 23 text extracts, which means that those proficient raters who read out text, read more than competent ones. Comment distribution is illustrated in Figure 11.7 and shows that raters turned to reading the texts much more often than to any other reading strategy.

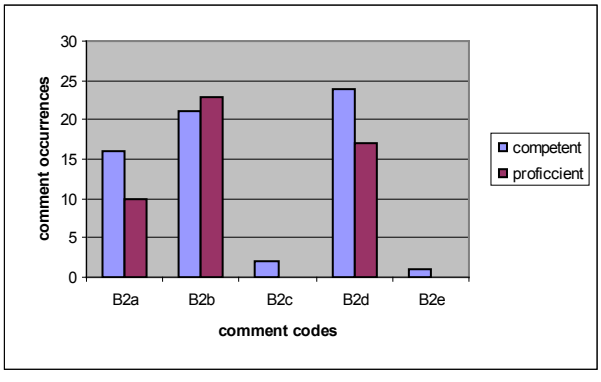


Figure 11.7. Reading comments when rating vocabulary of script N2

When providing examples to support rating decisions, ten (45%) competent raters read out 24 examples and eight (53%) proficient raters read out 17 examples. Although the same number of competent and proficient raters read out parts of text and one-word examples, they were not the same raters. There is no identifiable pattern in this respect: some raters read only text parts, while others read both parts and examples. Ten (45%) competent and four (27%) proficient raters employed neither comment type. The difference between reading out text parts and examples was, as described in Chapter Nine, to distinguish between raters' reading a word only or longer texts. When raters read out a word only, they concentrated on it regardless of context, but when they read out parts, they focused on context and evaluated them accordingly. There are examples for both in the extract in Excerpt 11.18 from a competent rater's protocol.

Excerpt 11.18: R1's reading comments when rating vocabulary of script N2

R1

TU	Rater talk
15	And there are words which don't belong to basic vocabulary like
16	*gorgeous*
17	and some other expressions *Say hi to everyone at work and give my love to your family* there are sentences which are difficult to understand
18	for example *I was thinking about you the other day when I realised that I haven't heard from you since I moved here*

However, none of the proficient raters summarised the script or read the rubric. This reading strategy was rarely used by competent raters; there were two occurrences of text summary and one of reading the rubric. The example in Excerpt 11.19 illustrates how a competent rater summarised the script when rating vocabulary.

Excerpt 11.19: R16's text summary when rating vocabulary of script N2

R16

TU	Rater talk
19	here is where, he says that life here is so different than back home but which is here, or what is here, what is home

The distribution of own focus comments was different for the two groups of raters: there was a total of 23 (mean 1.1) remarks by competent raters and 21 (mean 1.4) by proficient raters with different foci. Competent raters mentioned text comprehensibility three times, one expressed uncertainty, and four meditated on students' intention, as Excerpt 11.20 illustrates.

Excerpt 11.20: R11's remark on student's intention when rating vocabulary of script N2

R11

TU	Rater talk
11	But still there's motivation in the student to use these words and phrases appropriately

There were no comments in proficient raters' protocols when they were evaluating vocabulary. However, there were two comment types, which occurred

in proficient raters' protocols only: two comments on eligibility and one rater provided a solution for an error identified in the text.

Competent raters referred to students' proficiency five times, whereas proficient raters only once. Reflections on feelings occurred in seven comments by competent raters and in nine comments made by proficient raters. These comments in both groups express different feelings, but usually they were positive ones appreciating vocabulary choice. Errors were corrected in three cases by competent raters and in eight cases by proficient raters when evaluating vocabulary.

To sum up rating patterns of vocabulary evaluation, competent and proficient raters' focus was different: competent raters paid less attention to rating and reading strategies and they used fewer own focus comments than their proficient counterparts. Also, competent raters used fewer rating strategies than proficient raters who also turned to reading strategies more often. Raters paid attention to scale descriptors similarly; however, several raters provided evaluation of vocabulary in their own words. They made the most reading comments when reading the text and there was almost no attention paid to the task rubric when rating vocabulary.

The range of scores awarded to vocabulary was wide, especially competent raters' scores, as Figure 11.6 illustrates.

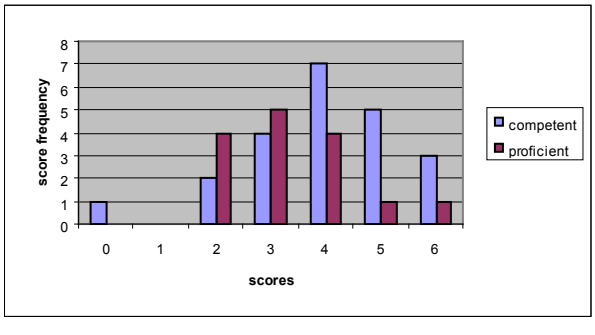


Figure 11.8. Scores on vocabulary of script N2

The benchmark for the rating criterion of vocabulary, as discussed above (see Table 11.8 for details) was 3 points. Proficient raters' scores were mostly one point below or above the score of 3 points. Competent raters, on the other hand, awarded higher scores; they mostly chose a score higher than the benchmark.

First, those comments are discussed that ended in the same scoring decision as the benchmark, and then some other decisions are examined. Raters who awarded three points to vocabulary referred to the range of vocabulary items

and remarked that the words were above the basic level (TU15), as the example from a competent rater's protocol in Excerpt 11.21 indicate.

Excerpt 11.21: R1's awarding the same score as the benchmark for vocabulary of script N2

R1

TU	Rater talk
1	**There are some expressions** and some sayings which are used here
2	For example *I felt like a fish out of water*
3	And there are words which don't belong to basic vocabulary

The score above the benchmark, which was a 4, was frequently chosen by competent raters, as they thought that features of vocabulary in script N2 corresponded to the scale descriptor and read it out (TU17), as it appears in Excerpt 11.22.

Excerpt 11.22: R12's reading the scale descriptor when rating vocabulary of script N2

R12

TU	Rater talk
15	**...**Vocabulary.
16	..is ...is ...okay, I think.
17	I think nice **expressions and appropriate** ...um nice expressions are used and appropriately
18	So...things like *how is it going,*
19	and *gorgeous*,
20	*furthermore*
21	So I would give ... a four to vocabulary.

The word "gorgeous" in the script appeared in many raters' comments, but they interpreted it differently. Rater R12 in her evaluation, as shown in Excerpt 11.22 cited the word (TU19) to exemplify that the text corresponded to the band descriptor for four points (TU21). Other raters referred to this word as an example for wide range of vocabulary, which was the descriptor for the top band, and they gave vocabulary six points. On the other hand, some raters noticed that the writer could not spell some words correctly and "gorgeous" was among them and awarded a score of two points, as the example from a proficient rater's protocol in Excerpt 11.23 illustrates.

Excerpt 11.23: RR13's noticing spelling errors when rating vocabulary of script N2

RR13	
TU	Rater talk
14	as far as vocabulary is concerned
15	I gave it a 2
16	Because there are problems with spelling
17	For example he cannot spell *was*
18	*gorgeous*,

These verbalised decision-making processes indicate that both competent and proficient raters had different ideas of what constituted wide and appropriate range of vocabulary, and what words belonged to basic vocabulary items, as they referred to the same word differently. In addition, they frequently referred to spelling, even if this criterion was not included in the scale.

To sum up, we could see that competent and proficient raters attended to different criteria and they used various strategies when they were rating vocabulary of script N2. Proficient raters paid more attention to this criterion, they made more comments with rating and reading focus, and they applied their own criteria. They chose a score from a narrower range of scores than competent raters.

1.1.4 Rating Grammar of Script N2

The third rating criterion in the rating scale was grammar and the two groups of raters' evaluation, as discussed above and illustrated in Table 11.1, was somewhat different, competent raters mean score was 3.95 (sd 1.4) and proficient raters' mean score was 3.4 (sd .99). The benchmark for grammar was a score of three points, which is highlighted in Table 14.10 to show the scores raters awarded and the differences between competent and proficient raters' choice of scores. More raters (15; 41%) chose a four, a score one point higher than the benchmark (3 points), which ten (27%) raters chose.

Table 11.10
Script N2 Grammar Score Frequencies (benchmark column highlighted)

Grammar score	0	1	2	3	4	5	6	Total
Frequencies (percentages)								
Competent	1 (5%)	0	2 (9%)	8 (36%)	7 (31%)	1 (5%)	3 (14%)	22
Proficient	0	0	4 (27%)	2 (13%)	8 (53%)	1 (7%)	0	15
total	1 (3%)	0	6 (16%)	10 (27%)	15 (41%)	2 (5%)	3 (8%)	37

The distribution of competent and proficient raters' scores, as it appears in Figure 11.9, shows different patterns: more competent raters chose the benchmark score and more proficient raters chose the score of four which was one point above it. The range of scores was wider in case of competent raters; they awarded scores from the whole range, whereas proficient raters chose scores between two and five points.

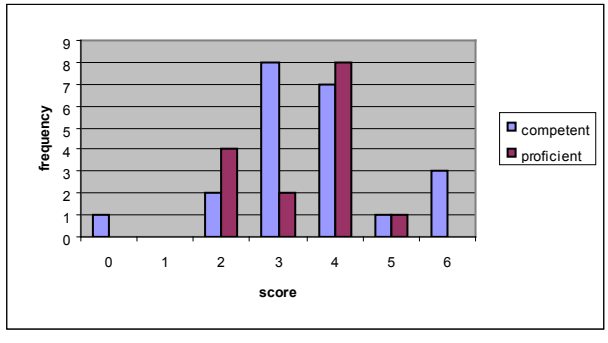


Figure 11.9. Score distribution for grammar of script N2

Competent raters mostly chose either 3 or 4 points and proficient raters mostly chose 4 points when they evaluated grammar. In what follows I provide a detailed analysis of competent and proficient raters' rating processes and of the way they arrived at a score.

Competent and proficient raters focused on rating strategies the most frequently, they made 115 (mean 3.1) rating comments, which was considerably more than reading and own focus comments: 72 (mean 1.9) and 34 (mean .9) comments, respectively (see Table 11.11 for details).

Table 11.11
Rating Foci for Grammar of Script N2

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	65 (3)	40 (1.8)	16 (.7)	121 (5.5)
Proficient	50 (3.3)	32 (2.1)	18 (1.2)	100 (6.7)
Total	115 (3.1)	72 (1.9)	34 (.9)	221 (6)

Competent raters' paid less attention to vocabulary than proficient raters, while the pattern of comment occurrences was similar to that of proficient

raters': they also mostly focused on rating remarks and the least on their own criteria, as Figure 11.10 illustrates.

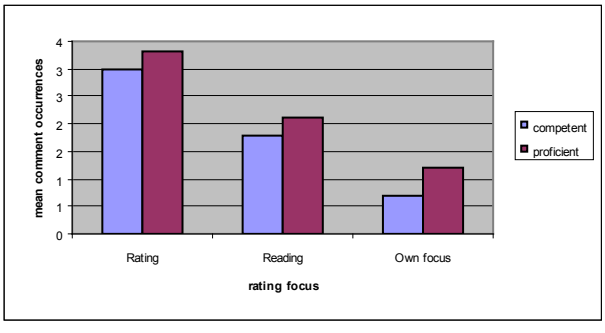


Figure 11.10. Rating foci when rating grammar of script N2

Raters paid the most attention to rating strategies when they were evaluating grammar of script N2. Their reference to the scale descriptors could be observed by looking at the comment occurrences in the first two subcategories in Table 11.12. According to the comment distribution, raters' attention to the two scale descriptors was similar (16 and 15 comments, respectively); however, competent raters paid more attention to accuracy (A3a) and made more comments (11) than on evaluation of structures (A3b) which was remarked six times. Proficient raters' attention was different, they paid less attention to accuracy (five comments) compared to structures (nine comments).

Table 11.12
Rating Comments for Evaluating Grammar of Script N2

	Codes											
Raters	A3a	A3b	A3c	A3d	A3e	A3f	A3g	A3h	A3i	A3j	A3k	Total (mean)
Competent	11	6	2	0	18	3	2	11	2	3	7	65 (3)
Proficient	5	9	12	3	12	0	0	3	0	3	3	50 (3.3)
Total	16	15	14	3	30	3	2	14	2	6	10	115 (3.1)

Comment distribution is illustrated in Figure 11.9 showing that proficient raters evaluated grammar in their own words considerably more often than competent raters.

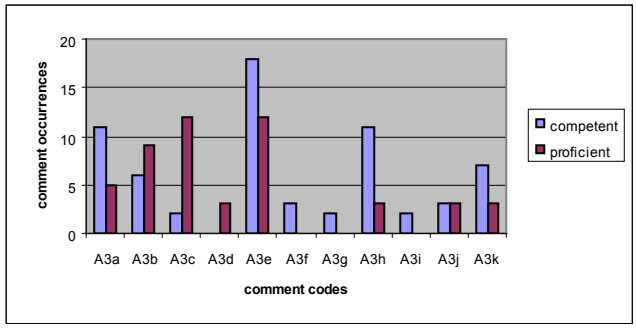


Figure 11.11. Comments when rating grammar of script N2

The example from proficient raters' protocols in Excerpt 11.24 illustrates how raters worded the evaluation of grammar in their own words.

Excerpt 11.24: RR15's rating grammar in own words of script N2

RR15

TU	Rater talk
9	he has very basic usage of grammar, very simple sentences

Three proficient raters added their own criteria: one of them, R3, referred to article use and RR3 and RR13 noted word order problems. Regarding score nomination, 18 competent raters out of 22 announced their score and 12 out of 15 proficient raters did the same. However, all scores were entered in the score sheets. None of the proficient raters added a reason for the score, revised their decision or referred to a lack of detail. Competent raters made few such remarks.

When raters noticed an error they either identified it (11 and 3 comments, respectively), or they corrected mistakes (2 and 5 comments, respectively). In addition, raters often switched to evaluation of vocabulary when dealing with grammar, as an example from a proficient rater's protocol in Excerpt 11.25 shows. RR8 elaborated on this confusion between the rating criterion of vocabulary and grammar.

Excerpt 11.25: RR8's switching focus when rating grammar of script N2

RR8

TU	Rater talk
26	for example *glass*
27	So, it seems, that *glass,* originally, ... not to mention that he wrote *two glass*
28	and not *two glasses*
29	But this is only one thing, and it doesn't count here, as we are looking at vocabulary now
30	seems that uses *glass* in the meaning of "bottle"

As mentioned above, raters sometimes switched focus, competent and proficient raters did so three times each in this respect. Score finalisation was not also used frequently; ten comments were made altogether in which raters finalised their score when rating grammar of script N2.

As far as reading focus is concerned, competent raters read less than proficient raters, as Table 11.13 indicates: competent raters 40 (mean 1.8) times, whereas proficient raters 32 (mean 2.1) times.

Table 11.13
Reading Comments When Rating Script N2

	Codes					
Raters	B3a	B3b	B3c	B3d	B3e	Total (mean)
Competent	20	16	0	4	0	40 (1.8)
Proficient	16	15	0	1	0	32 (2.1)
Total	36	31	0	5	0	72 (1.9)

Although five reading behaviour categories were identified, no rater summarised script content or read the rubric when evaluating grammar.

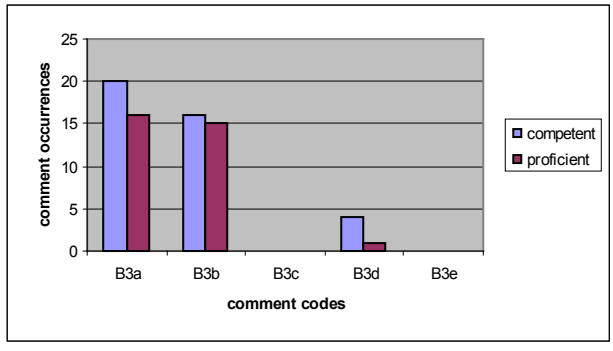


Figure 11.12. Reading comments when rating grammar of script N2

Competent and proficient raters mostly referred to reading the scale descriptors and reading out text parts when rating grammar and they rarely read out one-word examples, as is illustrated in Figure 11.12. Competent raters read the scale descriptors more frequently than they read text parts and made 20 and 16 comments, respectively. Proficient raters' comments were similar in this respect, they read the scale descriptors 16 times and text parts 15 times.

To sum up observations on raters' focus when they were evaluating grammar, proficient raters paid more attention to this criterion than competent raters did, they verbalised their evaluation in their own words more often and they seemed to be more concerned about structures than the accuracy of text.

Although both competent and proficient raters paid considerable attention to grammar and used different strategies for evaluation, their scores, as analysed above, show different patterns. Competent raters mostly chose a score of three or four, but three of them awarded the highest score, whereas more proficient raters chose a score of four more often than a three and none of them gave a score of six, but four raters gave a score of two.

First, those comments are examined which lead to the decision of awarding a score of three, which was the benchmark. A competent rater, as illustrated in Excerpt 11.26, said that she could not differentiate between grammar and vocabulary.

Excerpt 11.26: R11's score selection for grammar of script N2

R11

TU	Rater talk
13	Grammar
14	I would give three
15	Because there are inaccuracies
16	As I've already mentioned and
17	I cannot separate grammar from vocabulary

A proficient rater, on the other hand, referred to the writer's use of tense and connectives, and then made two remarks with his own focus before he chose the score, as illustrated in Excerpt 11.27.

Excerpt 11.27:R14's choice of score when rating grammar of script N2

R14

TU	Rater talk
33	The tenses ... regarding grammar, we could say that there are one or two connectives not in the right place
34	It could have been better
35	Let's say, maybe, he should deal with it a bit more, otherwise it would be a good composition
36	I would give it an average score, let's say a three

Another proficient rater mentioned the use of tenses as well, but in the end she chose a different score for grammar, as Excerpt 11.28 demonstrates.

Excerpt 11.28: RR18's choice of score when rating grammar of script N2

RR18

TU	Rater talk
23	Regarding grammar
24	He used Present Continuous, Past Continuous, Past Simple, and Present Perfect however not correctly. He used "would", so there were many tenses, but there are errors in tenses

At the same time a competent rater gave a score of 6 for grammar and evaluated the script differently saying that there were few inaccuracies and a variety in structures, as Excerpt 11.29 illustrates.

Excerpt 11.29: R18's choice of score when rating grammar of script N2

R18	
TU	Rater talk
14	For grammar
15	I gave 6 points
16	Because there were **only one or two inaccuracies**
17	And the structures are also ... rather varied

To sum up, raters seemed to interpret the features of grammar of the script differently, they often referred to the variety of tenses used by the writer and some of them considered whether structures were used correctly, while others did not.

11.1.5 Rating Organisation of Script N2

The last criterion in the rating scale was organisation and it was scored considerably higher by all raters than the benchmark. As Table 11.14 demonstrates, five raters (14%) who chose a score similar to the benchmark, one competent rater and four proficient raters. Others mainly chose scores higher than the benchmark, most competent raters awarded a score of three and quite a few awarded a score of two or four for organisation of script N2.

Table 11.14
Script N2 Organisation Scores (benchmark column highlighted)

Organisation score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Frequencies (percentages)								
Competent	1 (5%)	1 (5%)	6 (27%)	9 (41%)	3 (14%)	2 (9%)	0	22
Proficient	0	4 (27%)	4 (27%)	4 (27%)	3 (20%)	0	0	15
Total	1 (3%)	5 (14%)	10 (27%)	13 (35%)	6 (16%)	2 (5%)	0	37

The distribution of scores, as it appears in Figure 11.11, shows that four proficient raters chose either a one or a two or a three, and three chose four 4 points for organisation. Competent raters, on the other hand, chose mainly a three but their scores covered a wider range.

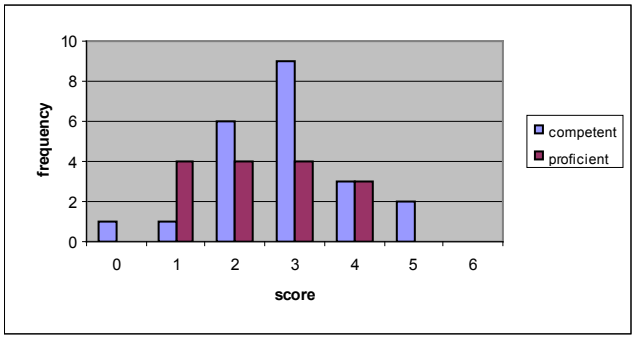


Figure 11.13. Score distribution for organisation of script N2

The following parts attempt to shed more light on raters' rating processes by examining their verbalised evaluations when rating organisation of script N2.

When dealing with organisation, raters applied the most rating strategies: 111 (mean 3) comments (see Table 11.15 for details). They made much fewer reading behaviour related comments: 66 (mean 1.8), and even fewer with their own focus: 19 (mean .9).

Table 11.15
Raters' Focus When Rating Organisation of Script N2

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	61 (2.8)	36 (1.6)	6 (.4)	103 (4.7)
Proficient	50 (3.3)	30 (2)	13 (.9)	93 (6.2)
Total	111 (3)	66 (1.8)	19 (.5)	196 (5.3)

Looking at the mean comment distribution, as Figure 11.14 illustrates, competent raters paid less attention to all three categories than proficient raters. It is also apparent that the number of proficient raters' comments with their own focus was considerably higher.

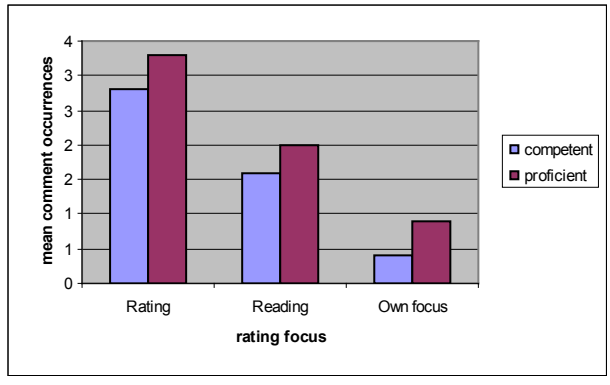


Figure 14.14. Rating foci for organisation of script N2

Raters did not comment on all subcategories, as shown in Table 11. 16, they did not revise their decisions and they never identified errors in texts. In addition, proficient raters never added a reason for choosing the score and only one of them added new criterion or referred to lack of detail. Competent raters rarely used these latter behaviour comments. Competent and proficient raters hardly ever changed focus, they made three comments altogether on change. Regarding other comments, there were some observable differences between the two groups of raters.

Table 11.16
Comments When Rating Organisation of Script N2

	Codes											
Raters	A4a	A4b	A4c	A4d	A4e	A4f	A4g	A4h	A4i	A4j	A4k	Total (mean)
Competent	8	9	6	4	19	6	0	0	2	1	6	61 (2.8)
Proficient	6	9	11	1	13	0	0	0	1	2	7	50 (3.3)
Total	14	18	17	5	32	6	0	0	3	3	13	111 (3)

Competent and proficient raters compared scripts to the first two scale descriptors similarly, as Figure 11.15 demonstrates; however, there were somewhat more comments made on the second descriptor, especially by proficient raters, which was evaluating links in texts.

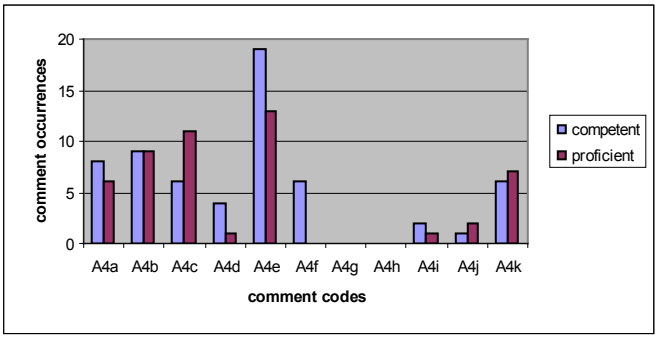


Figure 11.15. Comments when rating organisation of script N2

Comment occurrences show that raters used their own words in evaluating organisation of script N2, especially proficient raters turned to this strategy and they verbalised their evaluation in own words, as the example in Excerpt 11.30 shows.

Excerpt 11.30: R3's using own words when evaluating organisation of script N2

R3

TU	Rater talk
26	This piece of writing is somewhere halfway between a letter and a draft. He fails to structure his thoughts into paragraphs.

Five comments added criteria to the existing ones, four of which, as the example from a competent rater's protocol illustrates (see Excerpt 11.31), referred to paragraphing and a competent rater said that the letter had a relevant style.

Excerpt 11.31: RR5's additional criterion when rating organisation of script N2

RR5

TU	Rater talk
51	there are no separate paragraphs, for instance at eh beginning of the letter

As in the case of all four rating criteria, competent and proficient raters mostly announced the score they chose (19 and 13 comments, respectively) and all scores were duly entered in the score sheets.

Regarding giving a reason for score choice, six competent raters justified their score, as one of them in Excerpt 11.32.

Excerpt 11.32: R4's adding reason for score when rating organisation of script N2

R4	
TU	Rater talk
33	because there is some paragraphing, but still, there are scribbles in it, some confusion
34	the text is not properly organised
35	and as I have already mentioned, I would give it three points because at the beginning it feels that the text is fluent and the bullet points are not just put one after the other, but it is fluent nicely, but the ending is confused totally

She justified her choice in two comments (TU33 and TU35) and she explained the reason for the score she awarded.

Altogether three comments were made on the lack of detail and changing focus; six competent and seven proficient raters finalised their scores at the end.

Raters mostly read the rating scale descriptors when they were evaluating organisation of script N2, they made a total of 40 comments altogether (see Table 11.17 for details).

Table 11.17
Reading Comments When Rating Organisation of Script N2

	Codes					
Raters	B4a	B4b	B4c	B4d	B4e	Total (mean)
Competent	27	0	1	8	0	36 (1.6)
Proficient	13	7	5	4	1	30 (2)
Total	40	7	6	12	1	66 (1.8)

As the data show, the least they attended to the rubric, one proficient rater did so. Neither competent nor proficient raters summarised the scripts often; there was one competent rater and five proficient raters who turned to this strategy when rating organisation.

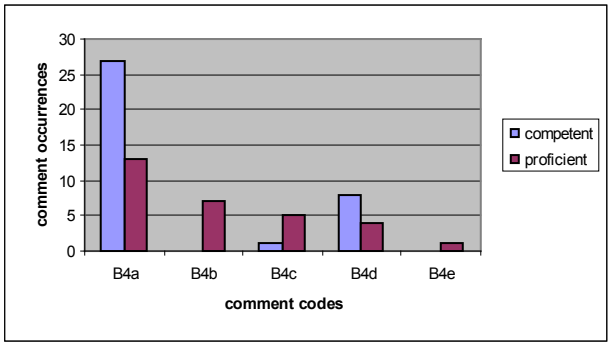


Figure 11.16. Reading comments when rating organisation of script N2

The distribution of reading comments, as it appears in Figure 11.16, represents different patterns, competent raters mostly referred to reading the rating scale and made 27 comments, while they did not read more words from texts but read out eight one-word examples. Proficient raters, on the other hand, made 13 comments related to reading the scale, they read texts seven times, summarised scripts five times, read one-word examples four times and there was one reference to the rubric.

Competent and proficient raters chose different scores when they were rating the organisation of script N2. Competent raters, as presented above, chose mostly a two or a three, whereas the proficient raters' choice was much wider, an equal number of raters chose a one, which was the benchmark, or a two or a three. Fewer raters chose a four when rating organisation.

A proficient rater who gave one point for organisation spent a lot of time and mentioned several different features of organisation before she chose the score, as Excerpt 11.33 illustrates.

Excerpt 11.33: R2's decision-making when rating organisation of script N2

R2

TU	Rater talk
1	There are many problems with organisation, that is with appearance
2	On the one hand we don't understand exactly what it is about
3	And there are … there are no separate paragraphs, a lot of sentences are just written in separate lines
4	There are a couple of sentences, 4-5, which are written in separate lines, so it is not transparent at all, why he wrote them the way he did
5	So as the content points are not elaborated well, it was impossible to connect them
6	So organisation is not good at all
7	I cannot see any logic in the letter
8	So organisation is 1 point

A competent rater, R13, arrived at a score of two after hesitating between a one and a two and referring to the layout and reading the descriptor for 4 points and another descriptor, for the score of two, as Excerpt 11.34 demonstrates.

Excerpt 11.34: R13's decision-making when rating organisation of script N2

R13

TU	Rater talk
31	**the layout reminds of a letter**
32	but not properly organised, **no clear logical link**
33	Hm … that's a one or a two
34	not properly organised
35	let it be a 2

As mentioned above, most competent raters awarded three points for organisation, as an example in Excerpt 11.35 shows. R1 first made a general remark, and then read out a descriptor from the scale for 4 points, then cited some examples before she awarded the score and she also made a remark on paragraphs.

Excerpt 11.35: R1's decision-making when rating organisation of script N2

R1

TU	Rater talk
40	The layout isn't good
41	**There are some links between the elements**
42	For example *furthermore*
43	*so*
44	*although*
45	*because*
46	*and*
47	but there are no paragraphs, just the sentences are written in a new line
48	I would give 3 points for organisation

Two raters gave five points for organisation, one of them, as presented in Excerpt 11.36, said that the elements were linked well and the text reminded her of a letter and finally decided on a score between the top score and a 4.

Excerpt 11.36 RR6's decision-making when rating organisation of script N2

RR6

TU	Rater talk
37	Elements are linked well, there is logical link between the elements
38	the letter has a style and ...
39	4 points by all means, maybe 5, it is a between a 6 and a 4
40	yes, organisation is 5 points

To sum up competent and proficient raters' rating processes when they were evaluating organisation, there was a considerable disagreement between the way raters perceived organisation. The range of scores was wide and especially proficient raters' scores were evenly distributed between a score of 1 point and 4 points. Although they seemed to pay equal attention to both layout and linking of scripts, they mostly focused on paragraphing. Raters rarely referred to reading strategies when dealing with organisation and their perception of scale descriptors was different from each other.

11.2 Ratings of Script N6: Benchmarks and Total Scores

The ten scripts for the present study, as mentioned in Chapter Seven were selected to represent a range of student performances including weak, medium and good performances. The scripts were chosen based on the benchmarks awarded by the researcher to have a point of reference for the study. According to the benchmarks, as Table 7.2 in Chapter Seven demonstrates, the best script out of the pool of ten scripts was script N6, which was awarded the top score in all four rating criteria in the rating scale. Competent and proficient raters' ratings were similar, as the total scores and rankings in Appendix 11.1 regarding competent raters and Appendix 11.2 regarding proficient raters show. Twelve (55%) competent raters ranked script N6 first; their next choice was script N10, which was ranked first by seven (32%) raters. Proficient raters chose one of the two scripts: nine (60%) proficient raters chose script N6 first and eight (53%) ranked script N10 first.

A comparison of benchmarks and mean scores of the two groups of raters on the four rating criteria, as illustrated in Table 11.18, shows that competent raters awarded lower scores than proficient raters to all four rating criteria.

Table 11.18
Benchmarks and Mean Scores on the Four Rating Criteria of Script N6

	Task achievement	**Vocabulary**	**Grammar**	**Organisation**	**Total**
Benchmarks	6	6	6	6	24
All raters' mean (sd)	5.79 (.47)	5.18 (.93)	5.26 (.83)	5.53 (.69)	21.76 (2.41)
Competent raters' mean (sd)	5.73 (.55)	5 (.98)	5.09 (.92)	5.5 (.8)	21.32 (2.68)
Proficient raters' mean (sd)	5.87 (.35)	5.4 (.83)	5.47 (.64)	5.53 (.52)	22.27 (1.91)

Competent raters awarded the highest scores to task achievement (5.73; sd .55) and the lowest scores to vocabulary (5; sd .98). The pattern is similar for the proficient raters, their mean score on task achievement was 5.87 (sd .35) and on vocabulary 5.4 (sd .83).

Table 11.19
Raters' Focus When Evaluating Script N6

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	231 (10.5)	172 (7.8)	91 (4.1)	494 (22.5)
Proficient	164 (11)	112 (7.5)	90 (6)	366 (24.4)
Total	395 (11)	284 (7.7)	181 (4.9)	860 (23.4)

Competent and proficient raters paid considerable attention to rating script N6, they made 395 (mean 11) rating, 284 (mean 7.7) reading-related comments, and 181 (mean 4.9) own focus remarks (see Table 11.19 for details).

Competent and proficient raters' mean comment distributions, as it appears in Figure 11.17, show different patterns. Competent raters remarked on rating fewer times than proficient ones, but they used somewhat more reading strategies than proficient raters.

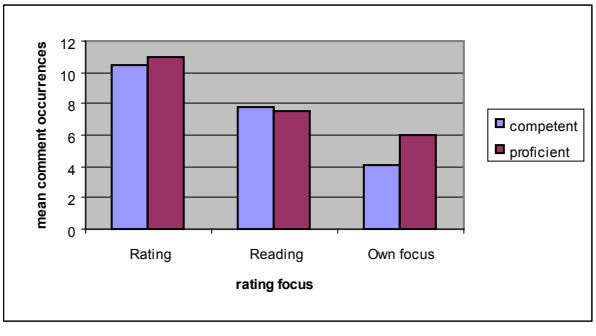


Figure 11.17. The three rating foci when evaluating script N6

As far as their own focus comments are concerned, proficient raters used them more often than competent raters. Rating processes on each rating are detailed in what follows.

11.2.1 Comments on Script N6 in the Pre-Scoring Stage

Rating written performance consists of several stages; the first is the pre-scoring stage, as discussed in Chapter Eight. However, not all raters made initial comments when they were rating script N6. Three (14%) out of 22 competent raters made such remarks and 6 (40%) proficient raters out of 15.

Raters in the pre-scoring stage referred to overall quality of the text, they mostly said that it was a well-written composition and sometimes stated that it was the best script they had read, as the examples from a proficient and a competent rater's protocol in Excerpt 11.37 show.

Excerpt 11.37.: R2's and RR14's remarks on text quality in the pre-scoring stage of script N6

R2

TU	Rater talk
1	It is perfectly written

RR14

TU	Rater talk
1	I found this one the best out of the ten I had to evaluate

There were one or two similar comments, except for a competent rater, R4, who made several different comments at the pre-scoring stage, sometimes referring to one or the other rating criterion, but it was only in her 17th comment that she announced the first rating criterion.

Six proficient raters expressed their appreciation of the script and they often made a personal comment, as an example in Excerpt 11.38 demonstrates.

Excerpt 11.38: RR1's comments in the pre-scoring stage when rating script N6

RR1

TU	Rater talk
1	after the first reading I have to tell that I was very satisfied with this letter, as everything is realised in it what a teacher can dream of in all levels

Other competent and proficient raters all started rating by announcing the first rating criterion in the rating scale, which was task achievement and they dealt with it without any initial remarks.

11.2.2 Rating Task Achievement of Script N6

As mentioned above, most raters started rating the scripts with the first rating criterion in the scale: task achievement. Thirty (81%) raters awarded the top score, which was the benchmark on task achievement of script N6, as Table 11.20 illustrates.

Table 11.20
Script N6 Task Achievement Score Frequencies (benchmark highlighted)

Task achievement score	0	1	2	3	4	5	6	Total
Frequencies (percentages)								
Competent	0	0	0	0	1 (5%)	4 (18%)	17 (77%)	22
Proficient	0	0	0	0	0	2 (13%)	13 (87%)	15
Total	0	0	0	0	1 (3%)	6 (16%)	30 (81%)	37

Competent raters chose a score from a wider range; a rater gave four points for task achievement of script N6. Other raters chose between a score of five or six points, but competent and proficient raters' score choice patterns were somewhat different, as Figure 11.18 demonstrates.

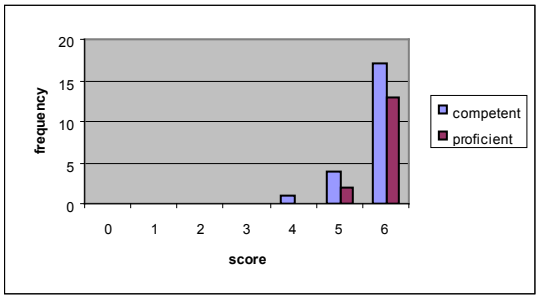


Figure 11.18 Task achievement scores of script N6

Fewer competent raters (17; 77%), chose the top score and more proficient raters (13; 87%) decided that task achievement of script N6 was worth six points.

When rating task achievement raters made a total of 211 (mean 5.7) comments: they attended to rating most often (87; mean 2.4 comments).

Table 11.21
Rating Foci When Rating Task Achievement of Script N6

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	53 (2.4)	59 (2.7)	24 (1.1)	136 (6.2)
Proficient	34 (2.3)	23 (1.5)	18 (1.2)	75 (5)
Total	87 (2.4)	82 (2.2)	42 (1.1)	211 (5.7)

They focused on reading less: they read 82 (mean 2.2) times and they paid the least attention to their own focus, 42 (mean 1.1) comments, as Table 11.21 shows.

The mean distribution patterns in Figure 11.19 illustrate the differences between raters' foci and the two groups of raters: they all paid the most attention to rating; competent raters read considerably more and own focus was the least and competent and proficient raters' attention was similar.

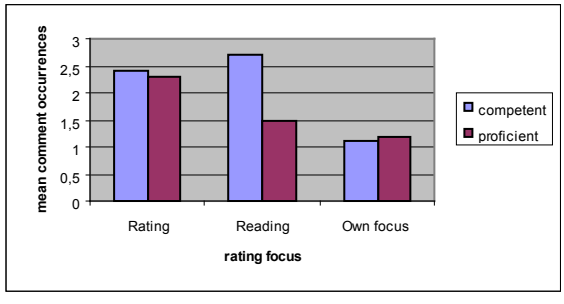


Figure 11.19. Comments when rating task achievement of script N6

As far as rating focus is concerned, Table 11.22 illustrates that there were two behaviour types that none of the raters referred to: they did not revise their decisions (A1g) and they never identified an error (A1h) in script N6. There were some other comments that raters rarely turned to.

Table 11.22
Rating Focus for Task Achievement of Script N6

	Codes											
Raters	A1a	A1b	A1c	A1d	A1e	A1f	A1g	A1h	A1i	A1j	A1k	Total (mean)
Competent	3	13	6	4	20	2	0	0	2	1	2	53 (2.4)
Proficient	1	3	6	5	14	1	0	0	0	2	2	34 (2.3)
Total	4	16	12	9	34	3	0	0	2	3	4	87 (2.4)

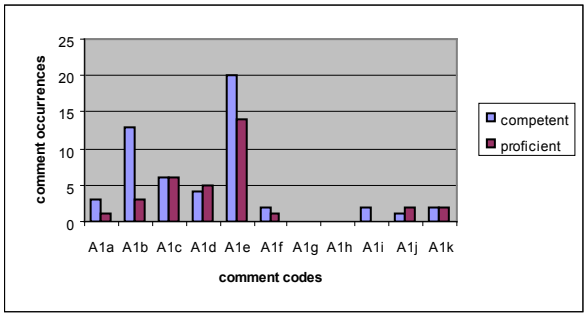


Figure 11.20. Comments when rating task achievement of script N6

As comment distribution in Figure 11.20 demonstrates, competent and proficient raters' attention to the rating scale descriptors was different. Although they considered both scale descriptors, competent raters seemed to deal more with the number of content points and made 13 comments, while they evaluated communicative goal three times. One proficient rater mentioned the first descriptor in the scale and there were three remarks achievement of communicative goal by proficient raters.

Raters sometimes articulated evaluation in their own words, as the example of a competent and a proficient rater in Excerpt 11.39 shows. A competent rater, R15, read out the scale descriptor (TU4) and then evaluated the script in her own words (TU5) and before returning to the scale descriptors again (TU6). The proficient rater, R8, followed a similar process, she also read the scale descriptor (TU6) and then verbalised her evaluation (TU7).

Excerpt 11.39: R15's and R8's evaluation of task achievement in own words of script N6

R15

TU	Rater talk
4	**so all content points are covered**
5	But ... and it is good ... is good and
6	**communicative goal is mostly achieved** yes and **all content points are covered**

R8

TU	Rater talk
6	**all five content points covered**
7	what needs to be included is there, so I think it is totally good, so

Nine raters, four competent and five proficient ones added their own criterion when rating task achievement. The four competent raters' referred to four different criteria: R18 mentioned creativity and colourfulness of the script, while there was another comment by RR2 on linking ideas and RR10 mentioned letter conventions. Proficient raters also noted transitions, letter conventions and they added the criterion of style. Similarly to rating other criteria both competent and proficient raters announced the score they chose with one or two exceptions.

There were three more rating strategies that were not used frequently: two competent raters pointed out lack of detail, as Excerpt 11.40 shows, they thought two ideas were not covered fully.

Excerpt 11.40: R1's and R7's remarks on lack of detail when rating task achievement of script N6

R1

TU	Rater talk
3	little is written about programmes he is planning

R7

TU	Rater talk
18	the other family members are not mentioned

Raters very rarely changed focus, there were three such comments altogether, a competent rater, R19, switched to grammar, and two proficient raters, one switched to organisation as he was talking about paragraphing and another, RR8, referred to good vocabulary use. I observed score finalisation four times.

To sum up raters' rating foci when evaluating task achievement, we can see that raters paid attention to rating scale descriptors, especially competent raters attended considerably more to content points and there were some raters who paraphrased their evaluation. They rarely used other strategies and there were no comments on revision of decision or error identification.

Regarding reading focus, raters mostly read out scale descriptors and text parts (see Table 11.23 for details). They sometimes summarised text content and there were 11 comments by competent raters with reference to the rubric. Except for two raters, they did not read out one-word examples from scripts.

Table 11.23
Reading Comments When Rating Task Achievement of Script N6

	Codes					
Raters	B1a	B1b	B1c	B1d	B1e	Total (mean)
Competent	21	17	9	1	11	59 (2.7)
Proficient	18	3	1	1	0	23 (1.5)
Total	39	20	10	2	11	82 (2.2)

Competent and proficient raters' comment occurrence distribution patterns were different, as shown in Figure 11.21 Competent raters read considerably more than proficient raters did and they summarised scripts, as well.

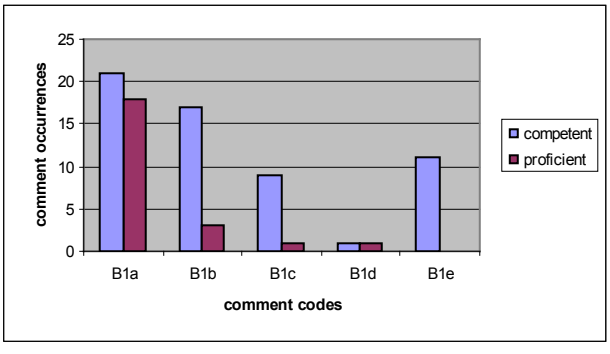


Figure 11.21 Reading comments when rating task achievement of script N6

Proficient raters, however, read mostly scale descriptors and they rarely read text parts or one-word examples. One proficient rater summarised the text and one read a one-word example from the text. Findings show that raters' reading foci were different, competent raters turned to different reading strategies more often than proficient raters.

Rating task achievement of script N6 resulted in agreement among the raters, as most of them chose the top score: six points. They commented on the content of the script positively, as an extract from a competent rater's protocol in Excerpt 11.41 shows.

Excerpt 11.41: R16's evaluation of task achievement of script N6

R16

TU	Rater talk
1	He communicated his intentions very successfully with a very good layout and however it's quite ambiguous what kind of time it was in his life in the sentence I had a time of my life.
2	He wanted it was the best time of my life I think.
3	I really liked how he introduced the topic of inviting Pat to Hungary.
4	So I gave 6 to task achievement,

The only competent rater who decided to give four points broke the script down into small units and rated them individually making reference either to task achievement or grammar, and made several comments with her own focus. Excerpt 11.42 is an extract from her protocol in which she evaluated task achievement and she read out a part (TU1) and then said it did not make sense (TU2) and gave another example (TU3) which, according to her, did not make sense either.

Excerpt 11.42 R7's evaluation of task achievement of script N6

R7

TU	Rater talk
1	He writes *Thank you … (I had a lovely time)*
2	He puts it into brackets as it doesn't make sense, I mean the last part
3	If we didn't take *I had the time of my life* into account
4	It doesn't make sense

11.2.3 Rating Vocabulary of Script N6

The second rating criterion in the scale was vocabulary. Competent and proficient raters mostly chose the same score as the benchmark: the top score of six points, as it is indicated in Table 11.24 Eighteen 18 (49%) raters chose it, eight (22%) awarded five points and ten (27%) gave a four for vocabulary. A competent rater assigned three points.

Table 11.24
Script N6 Vocabulary Score Frequencies (benchmark column highlighted)

Vocabulary score	0	1	2	3	4	5	6	Total
Frequencies (percentages)								
Competent	0	0	0	1 (5%)	7 (32%)	5 (23%)	9 (41%)	22
Proficient	0	0	0	0	3 (20%)	3 (20%)	9 (60%)	15
Total	0	0	0	1 (3%)	10 (27%)	8 (22%)	18 (49%)	37

Score distribution patterns, as Figure 11.22 shows, were different for competent and proficient raters. Competent raters chose a score from a wider range and there were considerably more, seven (32%), who awarded four, five (23%) gave five points and nine (41%) competent raters' score equalled the benchmark. Proficient raters mostly chose the top score, nine (60%) gave a six and other raters gave either a four or a five for vocabulary of script N6.

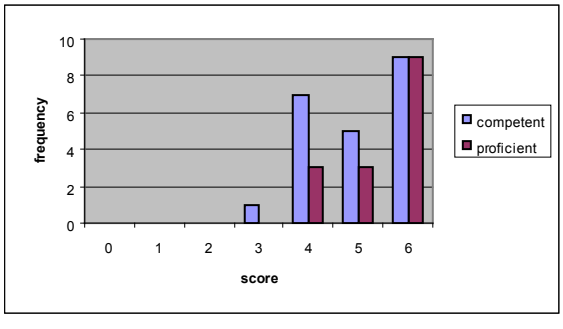


Figure 11.22. Distribution of vocabulary scores of script N6

Competent and proficient raters' focus when they were evaluating vocabulary of script N6 varied, as it appears in Table 11.25 They paid the most attention to rating and made 103 (mean 2.8) comments. They attended much less to reading, they read 78 (mean 2.1) times and there were few own focus comments (19; mean .5).

Table 11.25
Raters' Focus for Vocabulary of Script N6

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	60 (2.7)	36 (1.6)	6 (.3)	102 (4.6)
Proficient	43 (2.9)	42 (2.8)	13 (.9)	98 (6.5)
Total	103 (2.8)	78 (2.1)	19 (.5)	200 (5.4)

Competent and proficient raters' rating foci can be compared by looking at the means of comment occurrences, as Figure 11.23 shows, in which two different patterns can be observed.

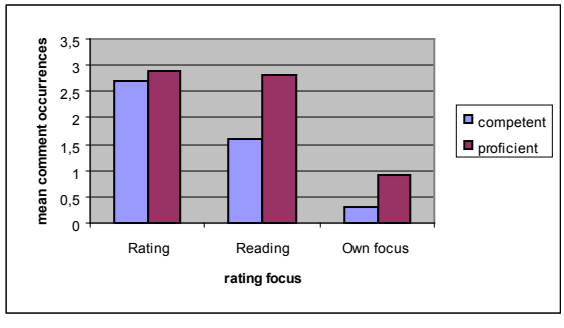


Figure 11.23 Rating foci when evaluating vocabulary of script N6

According to the means of comment occurrences, competent raters turned to rating strategies more often than to reading ones, but overall there were fewer comments, whereas proficient raters' rating and reading foci were similar to a certain extent and they made considerably more reading focus comments than competent raters. Proficient raters referred to their own foci more.

Rating focus comprised several comment types, as Table 11.26 illustrates. Raters mostly referred to scale descriptors (A2a and A2b), rated the script in their own words (A2c), added a criterion (A2d) or finalised the score (A2k). They sometimes added a reason (A2f) for the score and very rarely revised their decision (A2g), identified and error (A2h), referred to a lack of detail (A2i) or changed focus (A2j).

Table 11.26
Rating Comments for Vocabulary of Script N6

	Codes											
Raters	A2a	A2b	A2c	A2d	A2e	A2f	A2g	A2h	A2i	A2j	A2k	Total (mean)
Competent	7	5	8	4	20	4	2	1	1	2	6	60 (2.7)
Proficient	7	6	9	3	14	1	0	1	0	0	2	43 (2.9)
Total	14	11	17	7	34	5	2	2	1	2	8	103 (2.8)

The comment distribution patterns, as illustrated in Figure 11.24, reveal that competent and proficient raters paid similar attention to scale descriptors, they noted range seven times each and they remarked appropriacy five and six times, respectively.

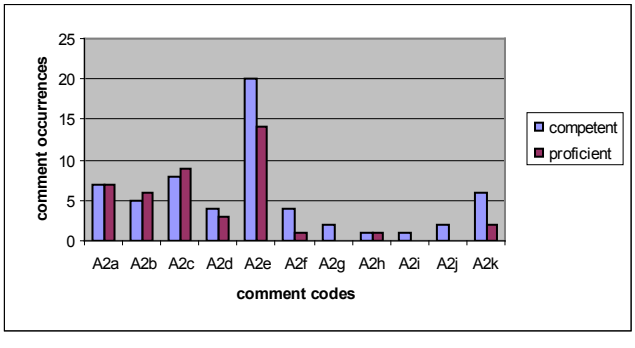


Figure 11.24 Rating comments for vocabulary of script N6

Regarding evaluation in their own words, both competent and proficient raters referred to this strategy, as the two examples, one of a competent rater and another of a proficient rater, illustrate in Excerpt 11.43.

Excerpt 11.43: R10's and R3's evaluation in own words when rating script N6

R10

TU	Rater talk
16	Ok, there are mistakes, but he uses...

R3

TU	Rater talk
10	Other expressions and words used in this letter highly correspond the task

There were seven own criteria that competent and proficient raters added when rating vocabulary. A competent rater, R18, mentioned authenticity and a proficient one, R14, mentioned that sentences were formed well, as Excerpt 11.44 demonstrates.

Excerpt 11.44: R18's and R14's own criteria when rating vocabulary of script N6

R18

TU	Rater talk
21	So it looks a bit authentic which is very good

R14

TU	Rater talk
17	And the sentences are put together nicely

Other competent raters, as RR9, RR17 and RR10 referred to correctness of spelling, while proficient raters, RR11 and RR13 added the same criterion on spelling. Regarding score nomination, almost all raters announced the score they chose when rating vocabulary of script N6.

Four competent and one proficient rater added a reason for the score and, as Excerpt 11.45 illustrates, remarked on writer's proficiency and vocabulary knowledge.

Excerpt 11.45. RR5's giving reason for the vocabulary score of script N6

RR5

TU	Rater talk
38	Because ... hm ... the letter reveals that the student's knowledge is at an intermediate level

There were only two comments by competent raters on revision of decision, one each on error identification and lack of detail. Two competent raters switched focus, turned to rating grammar and there were six occurrences of score finalisation. Proficient raters made much fewer comments in these categories;

they did not revise their decisions, did not mention lack of detail and did not change focus. One comment was made on an error and two on score finalisation.

As far as reading focus is concerned, we can see in Table 11.27 that raters mostly read more words from text, they made 42 such comments. They read the scale 18 times and quoted 16 one-word examples. There were only two proficient raters who summarised the script when rating vocabulary of script N6. No raters turned to the rubric when rating vocabulary.

Table 11.27
Reading Comments When Rating Vocabulary of Script N6

	Codes					
Raters	B2a	B2b	B2c	B2d	B2e	Total (mean)
Competent	12	15	0	9	0	36 (1.6)
Proficient	6	27	2	7	0	42 (2.8)
Total	18	42	2	16	0	78 (2.1)

The comment occurrence distribution, as Figure 11.25 illustrates, was different for competent and proficient raters. Competent raters read the scale and texts to a similar extent (12 and 15 comments, respectively), while proficient raters read the scale considerably fewer times (six) than more words from the text (27). Proficient raters read one-word examples from the script more often than they read the scale (seven and six comments, respectively).

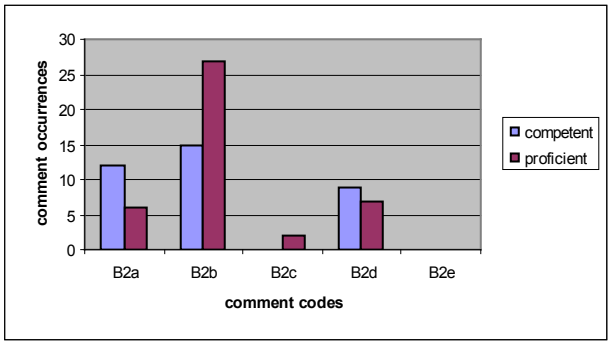


Figure 11.25 Reading comments when rating vocabulary of script N6

To sum up rating processes when dealing with script N6, raters mostly attended to rating criteria by comparing the script to the scale and verbalising evaluation in their own words. Competent and proficient raters rarely turned to

other strategies. Reading focus seems to be different, proficient raters focused more on text than competent raters did.

As far as scores are concerned, as described above, most raters chose the top score; however, many raters chose four or five points. The examples of raters' decision-making processes reveal how they arrived at a score.

First, I examine the top score, which was the benchmark. Nine raters in both groups chose six points and they arrived at the decision similarly: they read out the scale descriptor and read the script, as an example from a competent rater's protocol demonstrates in Excerpt 11.46

Excerpt 11.46.10: RR2's awarding the same score as the benchmark for vocabulary of script N6

RR2

TU	Rater talk
16	Vocabulary
17	*... bright ideas*
18	*great opportunity*
19	so that's quite a **wide range of appropriate words and expressions**
20	so that's a 6

Raters used similar strategies for their decision of lower scores, they chose a scale descriptor from a different band and they often mentioned mistakes, as Excerpt 11.47 from a competent rater's protocol shows.

Excerpt 11.47: R1's rating processes of vocabulary of script N6

R1

TU	Rater talk
6	Vocabulary
7	**Good and appropriate range of words and expressions** used
8	But the expression *at the end*
9	Isn't used correctly in the sentence
10	*I didn't have bright ideas about what to get but at the end I only got her a shirt*
11	I gave 4 points for vocabulary

The only competent rater, R7, who gave three points for vocabulary did not provide any evidence for her choice, she nominated the criterion and announced the score.

11.2.4 Rating Grammar of Script N6

The third rating criterion in the rating scale was grammar. Competent and proficient raters' scores and the benchmark are shown in Table 11.28.

Table 11.28
Script N6 Grammar Score Frequencies (benchmark column highlighted)

Grammar score	0	1	2	3	4	5	6	Total
Frequencies								
Competent	0	0	0	1 (5%)	5 (23%)	7 (32%)	9 (41%)	22
Proficient	0	0	0	0	1 (7%)	6 (40%)	8 (53%)	15
Total	0	0	0	1 (3%)	6 (16%)	13 (35%)	17 (46%)	37

Raters chose a score for grammar from a wide range: a competent rater gave three points, six (16%) raters awarded four points, 13 (35%) raters gave five and the majority, 17 (46%) gave the top score, six points.

Competent raters, as Figure 11.26 illustrates, apart from one rater, who gave three points, mostly chose a high score: nine (41%) raters gave six points, seven (32%) awarded five points and there were five (23%) raters who chose four points.

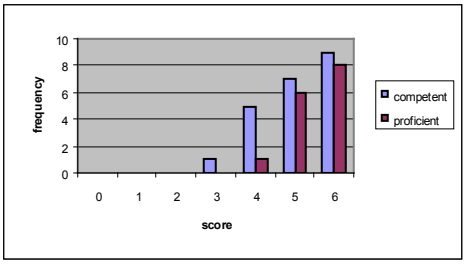


Figure 11.26 Grammar scores of script N6

Proficient raters, on the other hand, mostly gave six points: eight (53%) of them gave the top score, six (40%) awarded five points and one rater gave four points. Observations on the rating processes and the way raters arrived at a score are presented in the following sections.

Raters' focus when evaluating grammar centred on rating strategies: they made 112 (mean 3) rating comments (see Table 11.29 for details). However, their reading focus was also considerable, 80 (mean 2.2) comments were made with

a reading focus. In addition, raters frequently used their own focus comments: 46 (mean 6.4) times.

Table 11.29
Rating Foci When Rating Grammar of script N6

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	65 (3)	49 (2.2)	25 (1.1)	139 (6.3)
Proficient	47 (3.1)	31 (2.1)	21 (1.4)	99 (6.6)
Total	112 (3)	80 (2.2)	46 (1.2)	238 (6.4)

Competent and proficient raters' comments were similar, as the mean comment distribution patterns in Figure 11.27 show.

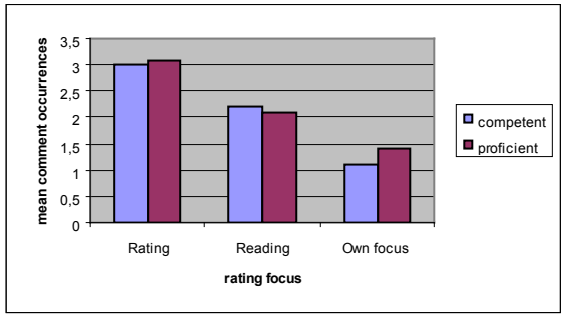


Figure 11.27 Rating foci when rating grammar of script N6

The raters focused on rating most often and competent raters made fewer comments in this respect. Reading focus was similar; however, competent raters made somewhat more reading remarks than proficient raters. The biggest difference was observed between own focus comments, competent raters made fewer than proficient raters. The comments in the eleven rating focus subcategories are in Table 11.30 to illustrate what comment types were used by raters when rating grammar of script N6. Raters attended to the rating criteria either by comparing the script to the scale descriptors (A3a and A3b) and made 21 and 11 comments, respectively on either descriptor or paraphrased the criteria 19 times.

Table 11.30
Rating Comments When Evaluating Grammar of Script N6

	Codes											
Raters	A3a	A3b	A3c	A3d	A3e	A3f	A3g	A3h	A3i	A3j	A3k	Total (mean)
Competent	12	7	9	0	20	3	2	8	1	0	3	65 (3)
Proficient	9	4	10	1	13	1	2	2	0	0	5	47 (3)
Total	21	11	19	1	33	4	4	10	1	0	8	112 (3)

The distribution of comments, as it appears in Figure 11.28 shows that both competent and proficient raters paid more attention to the first descriptor in the rating scale and commented on accuracy more than on structures.

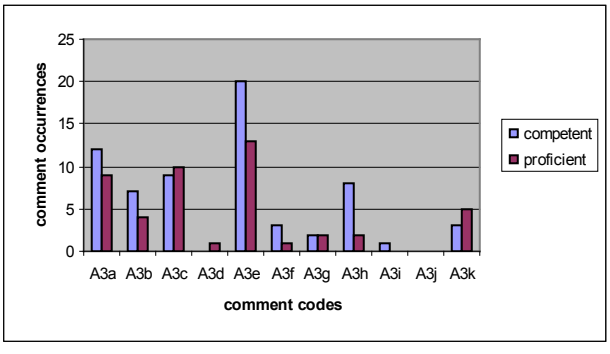


Figure 11.28 Rating comments for grammar of script N6

Both competent and proficient raters evaluated the rating criterion in their own words, as an example of a competent rater and a proficient one in Excerpt 11.48 illustrates.

Excerpt 11.48: R10's and R6's using own words when rating grammar of script N6

R10

TU	Rater talk
18	and grammar
19	is also pretty good

R6

TU	Rater talk
12	he has used grammar properly, so I don't have any problems with it

A proficient rater, RR12, added an own criterion when she was rating grammar: the use of inversion by the writer. Nearly all raters announced scores, there were only four who failed to nominate the score; however, they thoroughly entered all scores into the score sheet. Three competent raters and one proficient one justified their choice of score, as Excerpt 11.49 demonstrates.

Excerpt 11.49: R5's and R8's score justification when rating grammar of script N6

R5

TU	Rater talk
20	there are only small mistakes that's why I would give it a 5

R8

TU	Rater talk
40	So, apart from this there are no serious mistakes, and the small mistakes, how many, one, two and here listing at the end, it adds up to three and two words are misspelled

On two occasions raters revised their decision and chose a different score. Regarding error identification, there were eight remarks by competent and two by proficient raters. A competent rater, R7, commented on the lack of detail and said that an adjective was missing from one of the sentences. No raters changed focus when rating grammar of script N6 and three competent and five proficient raters finalised their scores.

There were 80 (2.2) reading behaviour related comments by raters (see Table 11.31 for details). They paid the most attention to the script and made 44 reading more words from the text comments and eight times, they read out one-word examples. There was only one competent rater who summarised the script and none of the raters read the task rubric.

Table 11.31
Reading Comments for Grammar of Script N6

	Codes					
Raters	B3a	B3b	B3c	B3d	B3e	Total (mean)
Competent	19	22	1	7	0	49 (2.2)
Proficient	8	22	0	1	0	31 (2.1)
Total	27	44	1	8	0	80 (2.2)

The comment distribution, as it appears in Figure 11.29, illustrates that competent raters read the scale considerably more frequently than proficient raters: 19 and eight times, respectively.

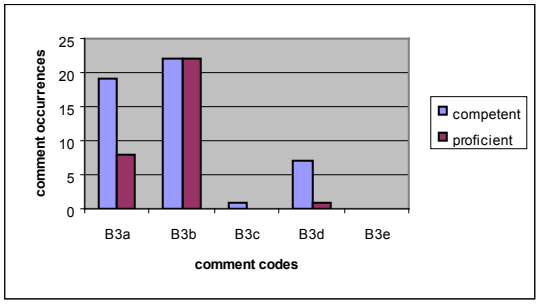


Figure 11.29 Reading comments when rating grammar of script N6

Both competent and proficient raters focused the most on the script when rating grammar, the difference was considerably more between proficient raters' reading the scale and the script, they attended much more to the script. In addition, they read out one-word examples, competent raters read out seven words, while a proficient rater one.

To sum up how raters attended to rating criteria, they compared the scale descriptors to the script and while doing so they focused on accuracy more than on range of structures. In several comments they used their own words for evaluation. As far as reading focus is concerned, proficient raters focused more on the script, as they used considerably more reading the script remarks than competent raters, who also read much from the script, but at the same time they frequently read out the scale as well.

The scores competent and proficient raters awarded for grammar on script N6 were mainly either five or six points. Competent raters chose a score from a somewhat wider range, four of them gave four points and one competent rater awarded three. Proficient raters' scores were less varied: one chose a score of four; all the others chose either five or six points. First, those rating processes are examined which resulted in raters' decision similar to the benchmark (six).

Competent raters most often read out descriptors from the scale and gave some examples before awarding six points for grammar. Some competent raters made one or two remarks, as the two examples in Excerpt 11.50 illustrate.

Excerpt 11.50 R17's and R7's awarding the score similar to the benchmark of script N6

R17

TU	Rater talk
1	I would award this composition a 6 in each aspect
5	the letter is grammatically correct, however there are some minor mistakes

RR17

TU	Rater talk
14	I could not find any ungrammatical structures
15	So grammar is 6, too

Proficient raters' rating processes were similar to the way competent raters arrived at a score, in addition, some of them listed the structures the writer used in his composition, as Excerpt 11.51 demonstrates.

Excerpt 11.51: RR11's rating grammar of script N6

RR11

TU	Rater talk
1	because he uses wide **variety of structures**
2	he uses both Simple past, Simple Present, besides conditional, correct ending

Those raters, who awarded five points for grammar, mentioned the number of inaccuracies in the text as a reason for not giving the top score. A proficient rater read out the scale descriptor for the top score and still awarded 5 points (see Excerpt 11.52).

Excerpt 11.52: RR1's decision on the score for grammar of script N6

RR1

TU	Rater talk
14	**There are one or two inaccuracies**
15	That is why I would give it a 5

The only rater who gave a score of 3 points was a competent one, R7, whose rating process was uneven, she made frequent jumps from one criterion to another and provided several new focus comments, as an extract from her protocol in Excerpt 11.53 illustrates. She dealt with grammar three times, on two

occasions she interrupted her evaluation of task achievement with comments on grammar and at the end when she announced the score for grammar.

Excerpt 11.53: R7's rating process for grammar of script N6

R7	
TU	Rater talk
19	*My parents ... opportunity*
20	here he corrects the good version and write the word in brackets with double "p"
21	Later puts "cultures" in brackets and writes "countries" instead, although we do not know how many countries he visited. Later writes "culture"
22	there are mistakes quite a few in this letter

To sum up the way raters arrived at a score for grammar, the main concern in their decision was the number of grammar mistakes in the text, which influenced their choice of score.

11.2.5 Rating Organisation of Script N6

The last rating criterion in the analytic rating scale was organisation, for which most raters (22; 59%) gave the top score, as Table 11.32 indicates. More competent (14; 64%) than proficient raters (8; 53%) awarded six points (the benchmark). However, two competent raters chose a lower score: one gave three points, the other four when rating organisation of script N6.

Table 11.32
Script N6 Organisation Score Frequencies (benchmark column highlighted)

Organisation score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Frequencies (percentages)								
Competent	0	0	0	1 (5%)	1 (5%)	6 (27%)	14 (64%)	22
Proficient	0	0	0	0	0	7 (47%)	8 (53%)	15
Total	0	0	0	1 (%)	1 (3%)	13 (35%)	22 (59%)	37

Score distribution patterns were not identical for competent and proficient raters, as Figure 11.30 indicates; competent raters chose a score from a wider range; still, most of them gave the top score. Proficient raters' choice of score was almost identical, seven (47%) chose five and eight (53%) six points.

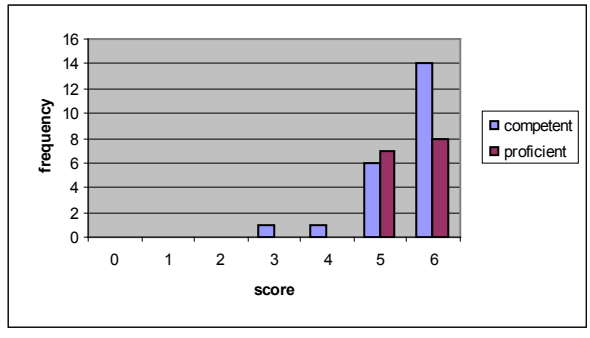


Figure 11.30 Score distribution when rating organisation of script N6

The following sections attempt to shed light on competent and proficient raters' rating processes and to examine how raters arrived at their scores for organisation.

Competent and proficient raters most frequently focused on rating strategies when they were evaluating organisation of script N6, they made 93 (mean 2.5) such comments, as illustrated in Table 11.33. They attended to reading much less frequently: 44 (mean 1.2) times. The fewest were their own focus comments: 13 (mean .4).

Table 11.33
Rating Foci When Rating Organisation of Script N6

Behaviour category	Rating (mean)	Reading (mean)	Own focus (mean)	Total (mean)
Competent	53 (2.4)	28 (1.3)	3 (.1)	84 (3.8)
Proficient	40 (2.7)	16 (1.1)	10 (.7)	66 (4.4)
Total	93 (2.5)	44 (1.2)	13 (.4)	150 (4.1)

The mean comment distribution patterns were different for competent and proficient raters: competent raters made fewer rating comments (53; mean 2.4) than proficient, who made 40 (mean 2.7), as visualised in Figure 11.31. Reading behaviour followed a different pattern: competent raters turned to reading strategies 28 (1.3) and proficient raters 16 (1.1) times. Own focus comments were rarely made by competent raters (three; mean .1), while proficient raters turned to them ten (mean .7) times.

Figure 11.31 Rating foci for organisation of script N6

Raters' rating processes were mostly characterised by their attention paid to the rating scale and rating criteria. Competent and proficient raters used other rating related comments very rarely, as comment occurrences in Table 11.34 indicate.

Table 11.34
Comments When Rating Organisation of Script N6

	Code											
Raters	A4a	A4b	A4c	A4d	A4e	A4f	A4g	A4h	A4i	A4j	A4k	Total (mean)
Competent	11	7	8	1	20	2	2	0	0	0	2	53 (2.4)
Proficient	6	8	5	2	14	0	0	0	1	1	3	40 (2.7)
Total	17	15	13	3	34	2	2	0	1	1	5	93 (2.5)

Competent raters attended more to the first scale descriptor on the layout of the script than to the logic of the text, whereas proficient raters paid more attention to text coherence and they compared the script to the rating scale (see Figure 11.32 for details).

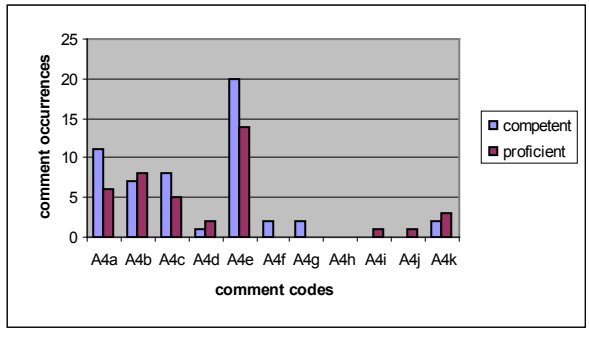


Figure 11.32 Comments when rating organisation of script N6

Raters frequently verbalised evaluation in their own words: they paraphrased criteria eight and five times, respectively. They mainly reflected on similar features of organisation and remarked on paragraphing, as the examples from a competent rater's and a proficient one's illustrate in Excerpt 11.54.

Excerpt 11.54: R16's and R6's using own words when rating organisation of script N6

R16

TU	Rater talk
1	and it's really organised, there are paragraphs

R6

TU	Rater talk
17	and he has used ... one, two, three, four, five ... six paragraphs to signal logical links

A competent and two proficient raters added their own criteria when rating organisation. The competent rater referred to writer's filling in space and the proficient rater referred to text division, as Excerpt 11.55 shows.

Excerpt 11.55: R13's and RR3's adding their own criteria when rating organisation of script N6

R13

TU	Rater talk
37	And it looks good, he filled in the space well

RR3

TU	Rater talk
14	And there is introduction, main part and conclusion

Raters announced the score they chose with three exceptions, but as it was the case with other rating criteria, there were no scores missing from the score sheets, raters failed to verbalise the score they awarded.

Regarding other rating related comments, there were only one or two of them: two competent raters added a reason and two revised their decision, while these comments did not occur in proficient raters' protocols. No rater identified errors in the script and only one proficient rater made a comment on lack of detail and change of focus. In addition, altogether five raters, two competent and three proficient ones finalised their scores.

Reading behaviour comment distribution, as it appears in Table 11.35 reveals that raters mostly read the scale descriptors (29 times) and sometimes more words from the script (seven times). However, six competent raters read one-word examples. Two proficient raters summarised the script content, but none read out individual words. Neither competent nor proficient raters read out the rubric, raters did not seem to pay attention to the rubric when they were rating organisation of script N6.

Table 11.35
Reading Comments When Rating Organisation of Script N6

	Codes					
Raters	B4a	B4b	B4c	B4d	B4e	Total (mean)
Competent	18	4	0	6	0	28 (1.3)
Proficient	11	3	2	0	0	16 (1.1)
Total	29	7	2	6	0	44 (1.2)

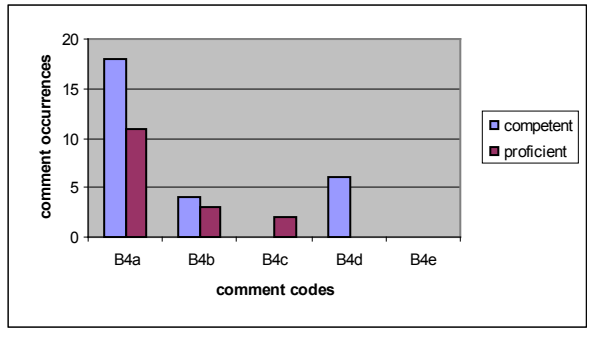


Figure 11.33 Reading comments when rating organisation of script N6

The comments were distributed differently: competent raters made considerably more rating comments than reading and, as mentioned above, six of them summarised the script, as it appears in Figure 11.33. Proficient raters turned to reading the scale and there was a similar number of comments related to reading more words from the script and summarising the script.

Raters mostly chose the top score, six points for organisation of script N6. However, two competent raters chose three and four points, respectively.

First, I examine those comments in which raters arrived at the top score from the scale. Fourteen competent and eight proficient raters awarded the score identical with the benchmark. Competent and proficient raters mainly read out both scale descriptors and announced the score of six points. An extract from a competent rater's protocol in Excerpt 11.56 demonstrates how.

Excerpt 11.56: RR17's evaluation of organisation of script N6

RR17

TU	Rater talk
21	because **the layout fully corresponds the task**
22	and **there is clear logical link between all text levels**

Looking at the two competent raters' ratings awarding lower scores than the others, we can see that one of them, R7, gave low scores (four, three and three points) for the other three criteria, and there was no evidence of her rating organisation, the score announced without any evaluation. The other competent rater, RR7, said that although the script reminded her of a letter, there was no link between all parts, as Excerpt 11.57 illustrates.

Excerpt 11.57: RR7's awarding a score of 4 for organisation of script N6

RR7	
TU	Rater talk
14	it corresponds the letter format
15	however, linking between parts is not always present
16	so I would give it a 4

In summary, competent and proficient raters evaluated organisation of script N6 using different patterns: competent raters turned to reading strategies more, while proficient raters commented more and they focused on rating strategies and used their own focus comments more often than competent raters did.

11.3 Conclusion

Competent and proficient raters paid considerable attention to different rating criteria when they were rating script N2. The script was chosen to exemplify a weak performance and was ranked last according to the benchmark and by most proficient raters. Competent raters' ranking was somewhat different: most of them ranked another script as last. Further disagreements between the two groups of raters were found in their rating processes. However, some variety was visible within competent and proficient raters' decisions as well.

Regarding the pre-scoring stage, more proficient raters made comments at this stage; their remarks were similar to those of competent raters. Raters' comments seemed to be contradictory; some thought the script was good and appropriate, whereas some considered it irrelevant.

Looking at rating processes when they were dealing with task achievement of script N2, all raters focused on content points and hardly any attention was paid to the achievement of communicative goal. This finding can lead to a tentative conclusion that raters found content points easier to evaluate than the achievement of the communicative goal. In addition, raters' perception of text content seemed to be different, as some of them, as their pre-scoring comments reflect, found it irrelevant and others felt that it conformed to task requirements. Competent raters used different strategies when evaluating task achievement: they turned to reading strategies much more often than proficient raters.

Rating vocabulary of script N2 resulted in different rating patterns for competent and proficient raters. Compared to rating foci when dealing with task achievement, although raters seemed to pay a similar amount of attention to rating and reading strategies and made a similar number of own focus comments, there were differences between competent and proficient raters'

comment distributions. The competent raters made fewer rating remarks than the proficient raters did, while proficient raters referred to reading strategies, especially reading more words from the text more often. Raters attended to scale descriptors, but they often evaluated the script in their own words. They obviously had problems with interpreting rating criteria, as what was considered wide range for one rater, was appropriate for another.

The next rating criterion was grammar. Competent and proficient raters used different rating processes: competent raters awarded either three or four points for grammar, while proficient raters mostly chose four points. Thus, competent raters were closer to the benchmark, which was three points, but they chose scores from a wider range. Competent raters commented less on grammar than proficient raters, who made more comments with a rating focus. In addition, proficient raters frequently evaluated grammar in their own words.

The rating criterion of organisation of script N2 resulted in the least agreement not only between the two groups of raters but within the groups, as well. Although competent raters' scores covered a somewhat wider range, proficient also chose scores from one to four. Raters rarely turned to reading strategies, they arrived at a score making mostly rating-related comments which were either comparing the script to the scale descriptor or evaluating it in own words. The only reading strategy they referred to, especially competent raters, was reading the scale descriptors, other reading related comments were infrequent. Findings show that even if raters used similar and less varied strategies for rating organisation, their decisions were different from each other as far as evaluation is concerned.

This chapter also provided an analysis of rating processes of the top script, N6, which was ranked first according to the benchmarks and by the majority of raters. As far as the awarded scores are concerned, proficient raters' scores were higher than competent raters' scores, thus their points were closer to the benchmarks.

Written performance assessment usually starts with a pre-scoring stage, which is characterised by some overall comments. However, not all raters commented on script N6 in the initial stage, more proficient than competent raters mentioned surface features or commented on the quality of the script.

Raters' focus could be characterised by substantial attention to rating and reading strategies and less to own focus comments. Raters made a similar number of rating and reading comments, but more proficient raters turned to their own focus comments when they were evaluating script N6.

Raters chose scores for task achievement from a narrow range; they awarded 5 or 6 points and made mostly rating related comments. Competent raters seemed to pay more attention to content points and all raters frequently paraphrased rating. Regarding reading, they mostly read the scale descriptors, but competent raters sometimes read the script and the rubrics. However, competent raters

made quite a few reading related comments, while proficient raters turned to reading much less often.

When rating vocabulary of script N2, raters chose scores from a wider range; especially competent raters' scores were varied: they awarded scores between 3 and 6 points. Raters' focus was different: competent raters made more rating related comments, while proficient raters paid similar attention to rating and reading. Rating comments were similar in kind, they compared the script to scale descriptors and they made several comments in which they rated the vocabulary in their own words. The distribution of reading related comments showed that competent raters read the scale descriptors and more words from the script similarly. However, proficient raters read the script considerably more frequently than competent raters and they attended more to the script than to the scale.

Grammar scores were chosen from a wide range similarly to scores for vocabulary, but the distribution was different: competent raters mostly chose a score of four, five or six points, whereas proficient raters mostly chose five or six points. Rating processes were characterised by considerable attention to rating strategies, which was similar for all raters, and they turned to reading less often. Proficient raters made more own focus comments than competent raters. Regarding raters' attention to rating criteria, they paid more attention to accuracy than to variety of structures and they often rated the criterion in their own words, especially proficient raters did so.

Rating organisation of the top script generated more comments from proficient raters than from competent ones, but their foci seemed to be similar, they made the most comments with rating focus and much fewer with reading focus. They mainly chose either 5 or 6 points, however, more competent than proficient raters chose the top score. Rating processes were characterised by remarks on rating criteria: competent raters paid more attention to layout features than proficient raters and there were remarks in which they used own words.

Raters' rating processes showed considerably more similarities than differences, especially when they were dealing with task achievement and organisation of the top script (N6), than rating processes for the weakest script (N2), where there was more disagreement between the two groups of raters and within the groups.

Chapter 12

Raters' Perception of the Rating Task and Thinking Aloud

Introduction

This study focuses on written performance assessment from the raters' perspective with the intention to reveal as much as possible about the way they arrive at a decision and their interpretation of scripts and the rating criteria. Raters in the study, as described in Chapter Seven, are novice raters with no previous experience in foreign language teaching and testing. They are pre-service English language teacher trainees in their last phase of teacher education. They took part in the rating exercise after they had received input in English language testing in general and testing written performance in particular, which was followed by rater training for the study. This chapter attempts to summarise the feedback they provided after accomplishing the rating task and to answer the fifth research question:

5. What is raters' feedback on the rating task?

First, I introduce the feedback sheet; then, I present raters' answers to the three questions in the order they appeared on the feedback sheet.

12.1 The Feedback Sheet

The feedback sheet included in the rating pack, (see Chapter Seven), asked the raters to comment on the course they had on language testing, the rater training and the rating task itself (see Appendix 7.6 for a copy of the feedback sheet). Raters' feedback data were collected by using an open-ended approach for two reasons: firstly, the intention was to avoid influencing and directing raters with specific questions. As all of them were novice raters and were not familiar with research in the area, I expected them to raise issues that could not be predicted in advance. Secondly, the rating task itself was time-consuming and demanding and I wanted them to concentrate more on the rating task.

Thirty-four feedback sheets were returned; all 22 competent raters and 12 out of 15 proficient raters provided their comments, whereas three raters did not fill in this data collection instrument.

Competent and proficient raters' comments on the three topics of the feedback sheet are analysed in the following sections. First, raters' remarks on the language testing course are presented; then, their opinion on the rater training for the rating task. Finally, I summarise their perception of the rating task.

12.1.1 Raters' Feedback on the Language-Testing Course

The one-semester elective seminar course on testing English as a foreign language (Chapter Seven) aims at familiarising pre-service English teacher trainees with main theoretical and practical issues in EFL testing. The first question on the feedback sheet asked raters for their opinion on the course.

Competent raters' feedback was mostly general, they summarised the course content and established links with their experience as language learners and their former studies in pedagogy. They also added that the course on language testing contributed to their expertise as would-be English teachers. They thought that the course raised their awareness of the importance of testing in foreign language education.

Competent raters found the course interesting, as an extract from a feedback in Excerpt 12.1 illustrates, they appreciated the balance between theory and practice.

Excerpt 12.1: R13's Feedback on the Testing in ELT Course

R13

I think the course was interesting; it was good to see the different task-types in testing. It was useful to do the tasks.

In addition, they mentioned that they had several opportunities to try out items. However, one competent rater thought that there was too much theory and thus some precious time was wasted instead of dealing more with practical issues of testing. The course made teacher trainees aware of the complexity and difficulties of test design. One of them remarked that she took part in seminars with enthusiasm and considered the tasks varied and creative.

Apart from the general comments, eight raters (36% of all competent and 24% of all raters), mentioned specific issues in their feedback. Two referred to their experience as language learners and related it to what they had learnt during the course, as one of them commented in her feedback (see Excerpt 12.2).

Excerpt 12.2: RR17's feedback on the effect of the course on testing

RR17

It [the course] also helped me personally to understand why exactly I did not like certain types of tasks and exercises during my skill practice.

Two competent raters reflected on the usefulness of watching the video recordings of oral examinations. Two other competent raters appreciated the handouts they got during the course. A competent rater regretted that they did not talk enough in seminars. Another rater said that she became aware of the importance of dealing with published materials cautiously, as they may contain mistakes.

Proficient raters' feedback was similar; however, more raters in this group mentioned a specific aspect of the course. Six proficient raters (50% of proficient and 18% of all raters), who highlighted an element of the course apart from the general remarks.

Three raters argued for making such a course obligatory for all teacher trainees, as according to them testing should be learned, as two extracts from their feedback in Excerpt 12.3 show.

Excerpt 12.3: RR14's and RR11's reaction on relevance of testing knowledge in teacher education

RR4

The course was very useful and I think the material should be obligatory for all students because it's necessary to learn not just how to teach but how to test as well.

RR11

The course was very useful and I think the material should be obligatory for all students because it's necessary to learn not just how to teach but how to test as well.

Similarly to some competent raters, a proficient rater, as Excerpt 12.4 illustrates, meditated on the difference between being a language learner and a language teacher. She commented on those activities in seminars when they had to do the tasks as if they had been test-takers and this experience made her think about becoming a teacher.

Excerpt 12.4: R2's opinion on the awareness of the difference between a language learner and a teacher

R2

It is also strange that suddenly we have to be on the other side; we are not students but teachers (or at least will be).

A proficient rater mentioned video watching, conducting oral examinations and listening tasks. Another rater pointed out the importance of dealing with faulty items to see the problems that might occur in item design and selection. These latter comments were similar to those, made by competent raters.

12.1.2 Raters' Feedback on Training for Rating Written Performance

The second question on the feedback sheet referred to the rater training conducted at the end of the seminar course on language testing. Training raters for written performance assessment is one of the most important elements of rating procedures. The nature and relevance of training raters for assessment of written performance is discussed in Chapter Four and rater training for the present study was designed bearing in mind the principles detailed there, as presented in Chapter Seven, which includes the detailed description of the training.

Twenty-two competent raters filled in the feedback sheet, but not all of them commented on all three questions. Three of them (14% of all competent and 8% of all raters) returned the sheets and did not write about the rater-training component of the language-testing course.

Nine (41%) competent raters made general remarks, summarised what the training was about, and mentioned that it gave them sufficient information about how to accomplish the rating task, as an extract from one of the competent raters' feedback in Excerpt 12.5 demonstrates.

Excerpt 12.5: RR2's feedback on the rater training with no specific focus

RR2

I think that the assessment training was very helpful and necessary, as most of us have not had any practice in assessing students' work. The detailed scale we were given made our job easier and ensured objectivity.

Ten (45%) competent raters mentioned one or two specific issues and some remarked on the length of the training they had.

The main concern for competent raters seemed to be the amount of practice they had during rater training. Although there was one rater, who said that they had enough time to prepare for the rating task, seven competent raters mentioned that they should have had more practice in both rating and thinking aloud. Three raters emphasised that they would have needed more practice in rating to gain more confidence, as the extract in Excerpt 12.6 exemplifies.

Excerpt 12.6: R11's comment on the amount of practice for the rating task

R11

The assessment task at the end of the course was good and as it was the main point of the course, I think we could have spent more time with it.

Competent raters expressed their opinion on practicing thinking aloud, which was the data collection method for the study. They found it difficult, and four of them remarked that they would have liked to see some examples of such protocols both taped and in writing, as they had never had to produce verbal protocol before. There is an extract from a rater's feedback in Excerpt 12.7 to illustrate what they thought of verbalising and then recording their rating processes.

Excerpt 12.7: RR10's comment on practicing how to produce a think-aloud protocol

RR10

I think there was little time left for preparing for the rating task. It would have been very useful for me if I had listened to a recorded evaluation, as I was not sure how to do it. I had never done such a thing before.

A competent rater remarked that the rating scale did not get enough attention in the training session and said that she could not justify scores of three and five. The other concern she raised was objectivity and although she acknowledged that the scale was easy to use, she thought that her evaluation of one aspect influenced the decision she made on another: Excerpt 12.8 demonstrates what she said.

Excerpt 12.8: RR5's concerns about objectivity of rating

RR5

> Another thing that I would like to mention concerns objectivity. Although the scale is easy to use and does not include too many details that would put a strain on the teacher's memory, sometimes I had the feeling that my assessment was not objective enough. I mean, is a student performed well from one aspect (e.g. vocabulary) I tended to give him/her good points further on.

Two (17%) of the twelve feedback sheets handed in by proficient raters did not contain comments on the rater training component of the language testing course. Proficient raters mainly provided details in their feedback on rater training; however, there were some general remarks as well, as Excerpt 12.9 shows. RR8 reflected in her remark that rater training raised teacher trainees' awareness in becoming a teacher.

Excerpt 12.9: RR8's comment on the effect the rater training had on becoming a teacher

RR8

> The training was appropriate and interesting. I really enjoyed assessing students, this time I could really feel as a teacher.

Regarding the amount of practice in rating and thinking aloud, proficient raters' opinion differed; two raters were satisfied with the amount of practice they had, one of them added that rating was a difficult task and would become easier next time, as the extract from her feedback sheet in Excerpt 12.10 exemplifies. She also refered to the logical structure of the training procedure.

Excerpt 12.10: RR4's feedback on the rater-training component of the testing course

RR4

> The training we had for the assessment task was clear, interesting and followed a step-by-step method. I've learnt how misleading 'impression marking' can be. Although the assessment scale is very clear and easy to use, sometimes it's still difficult to decide what points I should give. I'm sure, with practice it'll become easier.

Three proficient raters would have liked more practice in thinking aloud and one rater felt that the time allocated for rater training was not sufficient for the rating task, as her summary on rater training in Excerpt 12.11 illustrates.

Excerpt 12.11: R2's evaluation of the amount of time spent with rater training

R2

So, altogether, we should have spent a bit more time with training for the assessment task, but it was compensated with all the other things we learnt.

A proficient rater commented on the way he utilised the notes he made during the training and the training materials he received and said that he revised them before starting the rating task (see his words in Excerpt 12.12).

Excerpt 12.12: R14's comment on using the training materials during the rating task

R14

All the work we have done in the classroom helped me to accomplish the task. Before I started to work, I had read all the materials and I believe it helped me a lot.

These findings show that raters' awareness in testing writing performance was raised in rater training sessions, as a proficient rater summarised in her feedback what she had learnt (see Excerpt 12.13 for her comments).

Excerpt 12.13: R6's summary of the effect of rater training

R6

I thought that testing a piece of writing is always subjective and cannot be objective at all since it depends on the teacher's mood and feelings for his students. Now I can see that the teacher can be much more objective if he takes the trouble to use assessment scales. It is still difficult to be objective but at least they give some guidelines to assess writing, moreover, it can ensure inter-rater reliability.

She emphasised the change in her approach towards testing written performance and highlighted the need for using rating scales to ensure as much objectivity in judgements as possible.

12.1.3 Raters' Feedback on the Rating Task

Data collection for the present research centred on a rating task, comprising of rating ten scripts and verbalising evaluation. The verbalised evaluation had to be audio recorded and transcribed afterwards, as presented in Chapter Seven. The third point on the feedback sheet related to the rating task and intended

to find out how raters perceived it. All raters who filled in the feedback sheets reflected on the rating task. They mentioned both the rating processes and the think-aloud procedure.

Three competent raters (14%) summarised their experience on the rating task in general terms, as an extract in Excerpt 12.14 illustrates.

Excerpt 12.14: RR9's comments with general focus on the rating task

RR9

> The task was not easy, but it was interesting. It's hard to take into consideration the level of students. It was also hard to vocalise everything I think of.

The rater described the rating task in the example as an interesting one, another rater called it a particular experience, while another said it was a big challenge for her. The main concern seemed to be doing two things at the same time. Some competent raters remarked that concentrating on the rating task and verbalising thoughts turned out to be exhausting and demanding. The example from a rater's feedback in Excerpt 12.15 exemplifies how they perceived the pressure caused by rating and thinking aloud.

Excerpts 12.15: R9's comment on rating and thinking aloud at the same time

R9

> ... it was a tough job to concentrate on two things simultaneously. Assessing papers and paying attention to my thoughts at the same time is a very unusual situation. That's why the pace of my talk was a bit slower than usual.

A rater talked about transcription and said that although it was difficult and time-consuming to transcribe the audio recording, she found it useful and helpful, as she could notice some mistakes in her English speech. Effects of thinking aloud during rating was noted by other raters as well, apart of making mistakes and talking more slowly than usual, some raters mentioned frequent hesitations. One of the raters commented on thinking aloud similarly and acknowledged that verbalising her thoughts made her aware of what testing constitutes.

There were remarks on the rating procedure with special attention to the rating scale. A competent rater noted the difference between rating in the seminar and doing it on her own with no immediate help from either the trainer or fellow students. Some raters expressed their dissatisfaction with the scale and thought it did not always help, especially in cases, which did not seem to conform the

scale descriptors. There is an example in Excerpt 12.16 from a competent rater's feedback who expressed her concern about the scale similarly to other raters.

Excerpt 12.16: RR7's concerns related to using the rating scale

RR7

... the scale was a big help, although it does not help always, e.g. what to do with a student who wrote less than 150 words, but all the content points are covered? What to do if the student's handwriting is hardly legible?

Another rater said that the most difficult thing for her was to cope with the influence of the different criteria on one another. A competent rater considered herself to be a strict rater and noticed that she avoided the two extreme scores and tended to award scores from the middle bands (see an extract from her feedback in Excerpt 12.17).

Excerpt 12.17: R13's concerns about awarding scores according to the rating scale

R13

I tended to give average points: 3 and 4 (for me, average means 3 and I tended to forget that the maximum is not 5, but 6, and also point 3 is exactly on the halfway line of the chart). As I always try to avoid extremes, I seemed to avoid points 1 and 6.

A competent rater who found rating the last scripts easier than the first ones and she remarked that it would be interesting to rate the scripts again and compare the scores.

Similarly to competent raters, some proficient raters gave feedback with no specific focus and others commented on the rating processes and thinking aloud. Two (17% of all proficient raters) filled in the feedback sheets and provided general feedback on the rating task, as Excerpt 12.18 illustrates.

Excerpt 12.18: RR12's feedback on the rating task with no specific focus

RR12

The task itself was really interesting because I have never done similar task before. Moreover, I did not assess so many compositions earlier, not to mention the thinking aloud part of it. I think this assessment was a very useful help in our teacher training practice.

Those proficient raters who commented on thinking aloud while rating the scripts emphasised the complexity of thinking and rating at the same time and one of them remarked that he realised how difficult it was to verbalise thoughts. Another rater said that listening to her own voice was disappointing at first and she spent a long time transcribing the protocols, which sometimes made her lose patience. A proficient rater highlighted the complexity: R2 found the amount of talking surprising on one script, as Excerpt 12.19 illustrates.

Excerpt 12.19: R2's remark on thinking aloud during rating

R2

It is surprising though, how much one can talk or think about one single piece of writing. You have to consider everything, and in this the scale was a very great help. ... It was good for us as well that we had to listen to ourselves, not only because we could listen to our way of thinking, but also because we realised how long it takes to correct ten compositions properly.

In addition, she saw the benefits of thinking aloud while rating and said that she could learn something about her way of thinking and she became aware of the amount of time assessment takes. Other raters raised the issue of becoming aware of what constitutes written performance assessment as well, one of them said that at the beginning she felt rating very stressful but it became easier over time. Several proficient raters mentioned gaining experience, one of them mentioned it as a prerequisite of becoming an "objective" teacher, as an extract of her feedback in Excerpt 12.20 exemplifies.

Excerpt 12.20: RR16's comment on gaining experience in rating

RR16

In my opinion, becoming an objective teacher in assessment requires a lot of practice. Toward the end of the task I became faster and faster in recognising students' mistakes and errors, moreover, I found it easier to assess an essay.

Proficient raters mentioned several features of rating that they observed when assessing the scripts. One rater reflected on error gravity and said that sometimes it was not easy to distinguish between serious and less serious errors. Another remarked that although the rating scale was the best she had ever seen, she was not always sure if her scores were appropriate. Several proficient raters expressed this uncertainty; one of them said that she would probably give a different score some time later, and another expressed her worries related to

the script sequence. She noticed that the quality of individual scripts influenced her largely. In addition, another rater realised that she could not make a clear distinction between certain rating criteria and considered an aspect more than once, as the extract from her feedback in 12.21 illustrates.

Excerpt 12.21: R8's concerns about rating regarding criteria and time

R8

Sometimes I just couldn't decide whether to include a certain aspect in the assessment of grammar or vocabulary and so on. I'm afraid I sometimes considered a certain thing, such as spelling at least twice during the assessment of the same paper. I also realised that some hours later I would have given different points for the same script.

12.2 Conclusion

Raters' feedback on the course in testing EFL and their opinion on the rating task was generally positive; they acknowledged that the experience contributed to their expertise as future English language teachers.

As far as their feedback on the course in testing in ELT is concerned, raters mentioned the appropriate balance between theoretical input and practice, which made the course material both grounded in theory and closely related to practice. Apart from general remarks, there were some specific issues mentioned, especially proficient raters' feedback contained such comments. Raters mentioned the significance of a shift from being a language learner to a teacher and the importance of understanding the underlying principles of different tests. Competent raters appreciated watching video-recorded oral examinations and the handouts they received during the course. They frequently mentioned the value of examples, some of which were faulty items to make them aware what should be considered in item design and selection. In addition, proficient raters found the course of paramount relevance in foreign language teacher education.

Raters' feedback on the course in language testing was followed by comments on rater training. Some raters did not provide feedback on the training of rating written performance. Similarly to the comments on the testing course, some competent raters made general remarks on rater training, while some referred to specific issues. Several references related to the time devoted to practising rating and thinking aloud and raters felt that they did not spend enough time on them. Some competent raters expressed their concerns about using the rating scale. Proficient raters' feedback on rater training was similar; however, their opinion on the amount of training differed, they mostly thought it was sufficient.

Feedback on rater training showed that raters' awareness of the importance of training for written performance assessment was raised and they felt it necessary to deal with it as part of their teacher education.

Finally, raters' feedback on the rating task reflected their commitment to the task, they elaborated on the cognitive demand of rating and thinking aloud at the same time. They emphasised the time-consuming feature of the task, especially the transcription exercise. Raters found thinking aloud and transcribing their rating processes a rewarding experience and they thought that they promoted not only their awareness in expressing themselves in English, but their rating skills as well.

General Conclusions

Introduction

The study aimed to examine raters' decision-making processes during rating written performance. I carried out research with novice raters who had to evaluate ten scripts, which Hungarian learners of English had written. In order to make comparisons possible, the raters were divided into two groups based on their agreement with the benchmarks set by the researcher. The 37 participants were all novice raters with a similar background who went through the same rater training procedures; however, there were differences in their rating performances. Twenty-two raters' scores showed low correlations with the benchmarks and 15 raters' correlations were stronger with the benchmarks. Based on these relationships raters in the former group were labelled as competent and raters in the latter group as proficient raters throughout the study. The research into raters' rating processes was conducted using verbal protocol analysis. I collected the verbal data with think-aloud procedure to observe raters' decision-making processes. The investigation into raters' rating processes showed that although they tended to use similar processes, there were some differences as well.

In this final chapter, first I would like to summarise briefly the relevant literature, which provided a firm background to the research into written performance assessment. Then, I will highlight the main findings and draw conclusions in the order I posed the research questions. I also intend to summarise raters' feedback on the rating task. Next, I analyze the main theoretical frameworks of rating processes in the light of my findings. I employed verbal protocol analysis as the research method in the dissertation and I would like to elaborate on its value in looking into raters' thinking processes. Finally, I intend to point out at some implications and limitations of the study as well as propose ideas for further research into rating processes.

The motivation for the research was manifold: first, my interest in testing FL in general, and written performance assessment in particular had already generated some research in the area (Bukta, 2000; 2001; 2007; Bukta & Nikolov, 2002). Second, the data collection method of using verbal protocol analysis seemed to be extremely challenging. Third, several studies examine rater behaviour and researchers have compiled frameworks of rating processes in the international literature involving different raters, rating scales and performances, but they usually observe few raters (Cumming et al., 2002; Lumley, 2000; Milanovic et al., 1996; Wolfe, 1997). Thus, the observation of 37 novice raters and conducting the study in a Hungarian EFL context are the two main characteristics that make the dissertation innovative. In addition, the raters had to evaluate the same

ten scripts and I compared their ratings with the benchmarks. In what follows I would like to draw the main conclusions, provide some implications and explain some of the limitations of the results of my research into rater behaviour.

Background to Written Performance Assessment

The ability to produce written texts is a complex cognitive task and the processes in L1 and L2 show both similarities and differences (Weigle, 2002). The main difference is that in L2 writing the proficiency level of the target language plays a crucial role (Zamel, 1983). The models of communicative competence provide an insight into the features of language ability in general and writing ability in particular (Bachman, 1990). As writing is a skill closely related to education, I discussed some issues of language instruction with special attention to language ability assessment (Cohen, 1994b). Measuring language ability has been widely discussed recently and considerable research has been conducted into different areas including assessment of writing ability (Alderson & Banerjee, 2001; 2002). Writing ability assessment is strongly related to language performance assessment, in which the interaction between the different elements play a decisive role, as rating depends on the following characteristics: rater, scale, rating, performance, instrument and candidate (McNamara, 1996). Thus, the measurement error in written performance assessment can be attributed to any of these elements; that is why rating procedures should be carried out carefully (Alderson, et al., 1995). Research into rating processes resulted in different frameworks of the processes raters go through (Cumming et al., 2002; Lumley, 2000; 2002; Milanovic et al., 1996; Wolfe, 1996). The research carried out in this dissertation is based on findings of previous studies, which proved to be a firm background for investigating raters' rating processes.

Features of Rating Processes

The first research question aimed at finding out what features characterise competent and proficient raters' rating patterns. Results show that competent and proficient raters' rating patterns are mostly similar, although I identified some observable differences.

Regarding gender distribution, this study has not revealed any differences between female and male raters; the ratio of female and male raters in competent and proficient groups was similar.

Raters used L1, L2 or a combination of the two languages in their protocols, as there were no restrictions regarding language use. More competent raters used L1 exclusively and proficient raters switched between L1 and L2 which

may imply that raters' rating proficiency can be higher if they use the language in which the thought comes to their minds. This finding is supported by raters' feedback on the rating task in which they expressed some concern related to verbalising their thoughts in English during rating and listening to it afterwards. They found the task demanding and there were raters who discovered mistakes in their own talk in English.

As for the length of verbal protocols, earlier research pointed at individual differences in raters' verbosity and findings of the present study have confirmed this observation (Wolfe et al., 1998). Especially proficient raters' protocols showed a wide range in length, but they seemed to focus on each of the ten scripts more evenly than competent raters did. This finding may mean that they could pay equal attention to each script and fatigue did not influence their rating processes.

Sequencing was looked at from two aspects: to see whether raters changed the given order of the ten scripts, and to examine raters' sequencing of rating steps. The ten scripts were numbered from 1 to 10 and presented to raters in this order; still, five raters (14%) changed the order of scripts for unknown reasons. As focus of the study did not include the effect of sequencing, this change was ignored in the analysis.

Findings show that all raters most frequently attended to the four rating criteria (task achievement, vocabulary, grammar and organisation) as they appear in the rating scale in an orderly fashion; however, the way they rated scripts along the four rating criteria had different features. Raters, when rating any of the four rating criteria, developed their own schedule and employed several patterns as far as sequencing behaviour types are concerned, similarly to findings in earlier research (Lumley, 2000).

Raters tended to follow the rating stages suggested by the rating scales in an orderly way, as Lumley (2002) states, thus, findings of the present study confirmed this observation. Some raters made initial comments with no special focus at the pre-scoring stage in rating (Milanovic et al., 1996); especially proficient raters turned to such remarks and mentioned some features of appearance and content before they focused on the four rating criteria in the scale. Although Wolfe (1997) found that less proficient raters make more evaluative comments, my findings did not confirm this. I expressed some concern in connection with such comments in the pilot study, as they may threaten objectivity in marking (Cohen, 1994b), but findings of the main study did not confirm this point. There were more proficient than competent raters who included own focus related remarks in their ratings mainly in the pre-scoring stage of rating.

Observations of raters' systematic behaviour in sequencing rating stages showed individual differences. Although most raters followed the order of the four rating aspects in the rating scale, more competent raters jumped from one aspect to another, skipped some of them, or changed the order in the scale.

Regarding rating patterns for individual scripts, no single pattern was identified, even individual raters employed different patterns for different scripts. Nevertheless, proficient raters seemed to be more organised in this respect, as their sequences were more alike and included fewer jumps. However, rating the criterion of task achievement was different from rating vocabulary, grammar and organisation, as most raters, especially competent ones, relied on the script much more and considered the text as a whole or broke it up for evaluation. Competent raters used reading strategies considerably more frequently than proficient raters. Sequencing the different strategies during rating showed iterative patterns, no matter what order of strategies raters used, they often went back to a strategy when they were rating specific criteria.

Raters' Rating Foci

Raters' comments were collected in one of the four groups of management, rating, or reading strategies and own focus comments to investigate what raters focused on. Similar grouping of strategies appeared in Cumming et al.'s (2002) framework of rating processes. Regarding management strategies, all raters, as mentioned above, could follow largely the four rating criteria in the order they appeared in the rating scale. In addition, they could internalise the rating criteria, as they did not often comment on the rating process.

Although raters attended to all four rating criteria, their focus showed different patterns: their rating focus strategies were considerably more frequent when they were dealing with the aspect of grammar than when rating the other aspects. These findings confirm earlier research into performance assessment, which identified raters' considerable concern about linguistic features of texts (Cumming et al., 2002; Lumley, 2000; Kontráné, 2003). However, findings of the pilot study were somewhat different: raters in that study seemed to pay the most attention to rating the achievement of the communicative goal. The tentative conclusion might be that raters' focus can be influenced by the performance, as the texts in the pilot study were much shorter than in any other study and there were few linguistic features to evaluate.

Raters' reading strategy patterns showed differences: competent raters read the scripts much more frequently than proficient raters when dealing with task achievement. A similar pattern was found when raters were dealing with grammar: competent raters applied reading more frequently than proficient ones. Nevertheless, when rating vocabulary, proficient raters turned to reading strategies more often than competent raters did. Findings showed that the rating criterion of text organisation attracted the least reading focus from all raters: the reason may be that by the time they arrived at the fourth criterion in their rating process they had internalised the script. These findings imply

that proficient raters made their decisions with less reading when dealing with task achievement and grammar, but the rating of vocabulary needed more consideration of the script.

The number of comments with raters' own focus was similar; however, proficient raters referred to surface features, such as length and eligibility, more often and they frequently articulated their opinion on students' overall proficiency and possible intentions. These remarks appeared mainly at the pre-scoring stage of rating, indicating that proficient raters tended to refer to their own criteria.

Interpretation of the Rating Criteria

Raters, as discussed above, considered the four rating criteria in an orderly manner with some differences regarding rating foci. Their interpretation of the four rating criteria was characterized by both similarities and differences. Regarding raters' attention to the scale descriptors, we could see different patterns. When dealing with the aspect of task achievement, the two groups of raters were much more engaged in evaluating the number of content points covered than the achievement of the communicative goal. Rating the aspect of vocabulary showed a different pattern: competent raters attended less to the scale descriptors than proficient raters did. Yet another pattern occurred when rating grammar: competent raters attended more to accuracy and less to structure variety than proficient raters. Text organisation was also rated somewhat differently: competent raters paid more attention to layout features and less to links than proficient raters did. These findings show that raters did follow scale descriptors but they did not consider them equally, as the emerging patterns demonstrate.

Raters sometimes verbalised their evaluation of the criteria in their own words; they used this strategy much less frequently than when they compared the text to scale descriptors. Proficient raters evaluated the scripts in their words more often than competent raters did and this difference was especially significant when they were dealing with grammar and organisation. Regarding additional criteria, similarly to evaluation in own words, raters did not often refer to criteria not in the scale. If they did, it was mostly a detail of the criterion, such as remarks on appropriate ending of a letter, when evaluating organisation.

Raters' Script Interpretation

I examined raters' interpretation of the scripts by looking into the rating processes when they were dealing with the weakest (N2) and the top script (N6).

Findings show that raters found it much easier to identify the top script than the weakest one. Lumley (2000) chose those scripts for his research in which there was disagreement between the raters, as he claims that it is more difficult to rate misfitting scripts.

Regarding the weakest script, there were extreme evaluations; some raters thought the script irrelevant, while others considered it good. The majority, however, agreed that it was one of the weakest performances. Regarding the four rating criteria, some rating patterns were similar in the two groups of raters, whereas others varied. As observed earlier, raters mostly attended to the content points and they paid less attention to the achievement of the communicative goal when dealing with task achievement. This rater behaviour resulted in extreme perceptions of the script content; some raters concentrated on the number of content points covered regardless of the content. Competent raters' decision-making processes included much more reading than those of proficient raters. This pattern was the reverse for rating vocabulary; raters seemingly had problems with interpreting the descriptors: what some raters found wide range was appropriate for another in the same script. Competent raters' scores for the rating of grammar were closer to the benchmark and they paid less attention to this aspect than proficient raters did. Rating the aspect of text organisation resulted in the least agreement not only between the two groups of raters, but within the groups as well. Although raters' strategies were similar to some degree, the decisions were different.

Rating the top script (N6), on the other hand, revealed more agreement between the two groups of raters than in the case of weakest script. Competent and proficient raters rated task achievement and organisation similarly, they chose either five or six points and their rating processes were alike; however, competent raters read somewhat more when rating task achievement and proficient raters attended more to rating strategies when dealing with text organisation of script N6. As far as vocabulary is concerned, there was some disagreement between the two groups in their decisions. In addition, proficient raters' focus was different: they read more than competent raters did. The rating aspect of grammar generated similar remarks and raters' foci were alike, although proficient raters often evaluated the aspect in their own words and made more of their own focus comments.

Raters' Perceptions of the Rating Task

The rating pack for the rating task included a feedback sheet for the raters to comment on the input they had in testing in ELT, the rater training for the task and the rating task. The rating task was part of an elective seminar course on assessment in ELT. All raters found the proportion of theoretical and practical input

relevant, which provided an appropriate background for written performance assessment. They appreciated the amount of training for the task; however, some of them would have liked more practice and expressed their concern related to producing think-aloud protocols. The rating task raised their awareness in the importance of developing an expertise in performance assessment as a part of a teacher-training course. Finally, they pointed out the rewarding outcome of the rating task: they felt that listening to their evaluation and transcribing it contributed to the development of their teaching skills.

Placing Empirical Findings on the Rating Processes into Theoretical Frameworks

Research into assessment of written performance resulted in several frameworks that have been compiled using verbal protocol analysis as a data collection technique. Comparing the most recent frameworks and findings of my study in certain areas are in line with previous models, whereas in others outcomes suggest that models need to be modified. In what follows, I summarize these areas.

Rating is a problem-solving activity in which raters' behaviours are different: they can focus on the text and on the scale to varying degrees; in addition, rating is not a simple matching activity of the scale and the text (DeRemer, 1998). The scoring strategies that raters use depend on the rating task: they can be simple matching, scanning, evaluating, etc. (Greatorex & Suto, 2006). Raters' decisions are influenced by different characteristics, which stem from three sources: their interpretation of the rating criteria and the texts; in addition, their rating processes may differ. There is one feature that rating processes have in common: it is their recursiveness (Wolfe et al., 1998).

In the model proposed by Milanovic et al. (1996) the rating processes comprise four stages: pre-marking, scanning, quick reading and rating. Raters pay attention in these stages to different features and to varying degrees. Findings in the study on Hungarian raters confirmed that rating consists of different stages and that raters' attention shows variation. For example, although not all raters turned to commenting in the pre-scoring stage, mostly proficient raters did so.

Wolfe (1997) divided his raters into three groups according to their rating performance and analyzed their processes by comparing the three groups of raters. I employed a similar methodology for the observation of raters' behaviour: I compared their scores to the benchmarks and ended up with two groups of raters: competent and proficient ones depending on their agreement with the benchmarks. Wolfe's framework is built on the assumption that raters interpret the texts and rate them using the criteria that they have internalised. I found

that raters sometimes interpreted the same script differently and they did so with the criteria and ended up in various ratings. In addition, Wolfe found that competent raters made more jumps during rating and they made more personal remarks. My findings regarding jumps were similar, but I found that proficient raters turned to their own focus more often than competent raters did.

Lumley (2000; 2002) compiled the most detailed framework of the rating processes in which he identified three stages: reading, scoring and conclusion. These stages are included in the three levels of rating processes: instrumental in the centre of attention, institutional and interpretation levels. He examined raters' strategy use during rating and drew his model, which shows the complexity of rating processes. I identified somewhat similar strategies that raters used during rating, but I grouped them differently: I focused more on what he calls instrumental level and examined raters' attention to the scoring criteria and classified strategies into management, rating and own focus.

The model put forth by Cumming et al. (2002) identified two main strategy types: interpretation and judgement strategies and they examined them from three foci: self-monitoring, rhetorical and ideational, and language foci. Their division of strategies is similar to mine; however, I looked at strategies considering rating criteria.

Verbal Protocol Analysis as Data Collection Method

The study aimed to explore raters' thinking and focused on tracing their decision-making processes. Similarly to previous research (Cumming et al., 2002; Lumley, 2002; Wolfe et al., 1998), findings show verbal protocol analysis an appropriate way of following raters' thinking. Lumley mentions that rating in itself is a demanding task and if raters have to articulate their thinking, the task becomes even more stressful. However, thinking aloud also raises raters' consciousness in rating and they should concentrate more on the task (Lumley, 2000). My findings justify his statements, as several raters mentioned the cognitive load of the rating task and acknowledged their professional gain of the rating task in their feedback.

I found think-aloud procedure to be appropriate to look into how raters arrive at their decisions and the study revealed some important features of this data collection methodology. Regarding the data collection method, I made the following conclusions: first, data can get lost if the technical background for audio recording is insufficient and transcription is either difficult or impossible. Second, I found it useful that the transcription of data was done by participants; raters audio-recorded their rating processes and transcribed them afterwards.

It was beneficial for them, as they became more aware of the rating processes by listening and transcribing their thoughts. In addition, asking raters to provide transcriptions was practical and virtually no data got lost. Third, the coding scheme developed gradually bearing in mind the focus of the research and was checked for reliability, as described in the literature (Ericsson & Simon, 1993).

Implications of the Findings

Findings of the research can extend our perception of rating written performance and thus, complement rater training sessions and teacher training programmes by providing trainers with valuable information of raters' rating processes that can be integrated into training sessions. EFL teachers and teacher trainees could compare rating processes and develop their own strategies for rating written performance. They can be familiarised with different rating strategies and practice the use of rating scales. In addition, raters' focus on text characteristics shed light on what raters paid attention to. The study revealed that raters tended to concentrate more on the completion of content points than on the communicative value of texts. There fore, new studies can be conducted into finding reasons for the differences in raters' attention.

Finally, findings, as in earlier studies (Brown, Glasswell, & Harland, 2004; Kontráné, 2003), confirm that it is not expertise that plays a substantive role in rater variance. Novice raters with no previous experience in rating can reach agreement after training.

Limitations of the Study

Although 37 raters were involved in the inquiry and participants provided a huge database in their verbal protocols, it is difficult to make generalisations about rating processes in written performance assessment. The raters in this study were all novice raters and the rating task was not carried out in an operational setting of an examination centre. Thus, all conclusions refer to this particular population in a non-test situation and most probably, findings would be somewhat different with experienced raters in an operational rating situation. In addition, a single task was used in the research: no data on how raters' behaviour changes with task type and text type was collected. The same problem relates to the level of the texts. Writers of the texts were supposed to be on B1 level, therefore, the study did not explore different levels.

Raters' interpretation of vocabulary and structural characteristics needs more standardisation, what some raters appreciate, others may find erroneous. Deviation from the scale descriptors does not necessarily affect rating

performance, as findings showed that proficient raters made more comments with a general focus than competent ones, which can mean that internalisation of rating criteria plays a much more important role than direct attention to scale descriptors.

The data collection method most probably influenced participants; some raters reported difficulties when they were rating and thinking aloud at the same time and noted that the rating task with thinking aloud took considerable time. However, similar research conducted using verbal protocol analysis does not confirm this entirely, as Lumley, for example, says that although raters in his study found rating and articulating thoughts demanding, the results were not influenced by the method (2002).

Further Research

The dataset of the present study would lend itself to further investigations of rater behaviour; there are considerable differences among individual raters whose protocols could be examined employing case study method to draw up individual rater profiles.

In addition, raters' thinking can be analysed in more detail, the database can be exploited in more depth to see what differences and similarities can be found in scale descriptor and text interpretations. Raters' thinking processes could have been examined with follow-up questionnaires and interviews to enquire into raters' justifications of their decisions.

Raters' behaviour in written performance assessment can be looked at from the point of view of raters' strategy use. Attempts have been made to compare raters' behaviour to problem-solving (DeRemer, 1998) and decision-making (Greatorex & Suto, 2006), as they appear in psychology, to compile a taxonomy of rater strategies. This line of research could be extended by categorising raters' strategies and elaborating frameworks of strategy use in performance assessment.

Further research should be conducted in an operational setting to examine how raters' rating processes work in such a context. It would be informative for examination designers, rater trainers and teachers alike to see into the rating processes of the English writing tasks of the Hungarian school-leaving examination to gather more information on the examination that has been recently introduced.

References

Alderson, J. C. (1991a). Dis-sporting life. Response to Alastair Pollitt's paper:' Giving students a sporting chance: assessment by counting and by judging'. In J.C. Alderson, & B. North (Eds.), *Language testing in the 1990s* (pp. 60-70). London: McMillan.

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

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research. Selected papers from the 1990 Language Testing Research Colloquium* (pp. 46-58). Alexandria, Virginia: TESOL Publications.

Alderson, J. C. (1998). Testing and teaching: the dream and the reality. *novELTy, 5*(4), 23-46.

Alderson, J. C. (May, 1999). Testing is too important to be left to testers. Plenary address to the Third Annual Conference on Current Trends in English Language Testing. United Arab Emirates University, Al Ain and Zayed University, Dubai Campus.

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

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* 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., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*(3), 280-297.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching, 34*(4), 213-236.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(3), 79-113.

Archibald, A., & Jeffery, G. C. (2000). Second language acquisition and writing: a multi-disciplinary approach. *Learning and Instruction, 10*(1), 1-11.

Association of Language Testers in Europe (1998). *A multilingual glossary of language testing terms.* Cambridge: Cambridge University Press.

Bachman, L. F. (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. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F., Davidson, F., & Foulkes, J. (1993). A comparison of the abilities measured by the Cambridge and Educational Testing EFL test batteries. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research. Selected papers from the 1990 Language Testing Research Colloquium* (pp. 24-46). Alexandria, Virginia: TESOL Publications.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*(2), 127-150.

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

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Bárdos, J. (1988). *Nyelvtanítás: múlt és jelen.* Budapest: Magvető Kiadó.

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

Blattner, N. (1999). Demystifying writing assessment: empowering teachers and students. *Assessing Writing, 6*(2), 229-237.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: a review of the issues. *Language Testing, 15*(1), 45-85.

Brown, A. (2005). Self-assessment of writing in independent language programs: The value of annotated samples. *Assessing Writing, 10*(3), 174-191.

Brown, J. D. (2004). Performance assessment: existing literature and directions for research. *Second Language Studies, 22*(2), 91-139.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*(2), 105-121.

Brualdi, A. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research & Evaluation. 6*(2), Retrieved on October 3, 2000 from:

http://ericae.net/pare/getv.asp?v=6&n=2.

Bukta, K. (2000). Reflections on test-taking strategies of 7th and 11th grade Hungarian students of English. *novELTy, 7*(3), 48-59.

Bukta, K. (2001). Mit tanulnak a diákok angol órán? 7. és 11. évfolyamos tanulók angol nyelvtudásának vizsgálata. *Iskolakultúra, 11*(8), 36-48.

Bukta, K. (2007). Assessment of written performance: Tracing raters' decision making process. *Porta Linguarum, 8*(2), 21-41.

Bukta, K., & Nikolov, M. (2002). Nyelvtanítás és hasznos nyelvtudás: Az angol mint idegen nyelv. In B. Csapó (Ed.), *Az iskolai műveltség* (pp. 169-192). Budapest: Osiris.

Bukta, K., Gróf, Sz., & Sulyok, A. (2005). *7 próba érettségi angol nyelvből. Középszint.* Szeged: Maxim Kiadó.

Calkins, L., Montgomery, K., & Santman, D. (1999). Helping children master the tricks and avoid the traps of standardized tests. *ERIC Clearinghouse on assessment and evaluation.* Retrieved on July 18, 2006 from: http://www.eric.ed.gov/ERICWebPortal

Campbell, C. (1990). Writing with others' words: using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211-231). Cambridge: Cambridge University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5-35.

Chapelle, C., & Douglas, D. (1993). Foundations and directions for a new decade of language testing. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research. Selected papers from the 1990 Language Testing Research Colloquium* (1-24). Alexandria, Virginia: TESOL Publications.

Chapman, C. (1990). Authentic writing assessment. *Practical Assessment, Research & Evaluation. 2*(7). Retrieved on July 4, 2003 from:

http://www.ericae.net/pare/getvn.asp?v=2&n=7.

Chapman, D. W., & Snyder Jr, C. W. (2000). Can high stakes national testing improve instruction: re-examining conventional wisdom. *International Journal of Educational Development, 20*(6), 457-474.

Cho, Y. (2003). Assessing writing: Are we bound by one method? *Assessing Writing, 8*(3), 165-191.

Cohen, A., D. (1984). On taking language tests. *Language Testing, 1*(1), 70-81.

Cohen, A., D. (1989). Second language testing. In M. Celce-Murcia, & L. McIntosh (Eds.), *Teaching English as a second or foreign language* (pp. 486-506). New York: Newbury House.

Cohen, A., D. (1994a). Verbal reports on learning strategies. *TESOL Quarterly, 28*(4) 678-82.

Cohen, A., D. (1994b). *Assessing language ability in the classroom.* Boston: Heinle & Heinle.

Common European framework of reference for languages: Learning, teaching, assessment. (2001). Cambridge: Cambridge University Press.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Connor, U. (1999). Writing in a second language. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 306-310). Oxford: Elsevier Science ltd.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-65.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing, 18*(2), 207-24.

Cumming, A., & Riazi, A. (2000). Building models of adult second-language writing instruction. *Learning and Instruction, 10*(1), pp. 55-71.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 66-96.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107-45.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

DeRemer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

DeVincenzi, F. (1995). Language tests and ESL teaching. Examining standardized test content: Some advice for teachers. *TESOL Quarterly, 29*(1), 180-87.

Dickson, P., & Cumming, A. (Eds.) (1996). *Profiles of language education in 25 countries.* Slough: National Foundation for Educational research.

Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). What does research say about assessment? *NCRL, Oak Brook.* Retrieved on December 13, 2005 from:

http://www.ncrel.og/sdrs/areas/stw_esys/4assess.htm.

Djigunović, J. M. (2006). The role of affective factors in the development of productive skills. In M. Nikolov, & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 9-23). Pécs: Lingua Franca Csoport.

Douglas, D., & Chapelle, C. (Eds.) (1993). *A new decade of language testing research. Selected papers from the 1990 Language Testing Research Colloquium.* Alexandria, Virginia: TESOL Publications.

Douglas, D. (1995). Developments in language testing. *Annual Review of Applied Linguistics, 15*, 167-187.

Dörnyei, Z. (1988). Language testing. In Z. Dörnyei, *Psycholinguistic factors in foreign language learning* (pp. 8-46). Unpublished PhD dissertation. Budapest: Eötvös Lóránd University.

Eisterhold, J. C. (1990). Reading-writing connections: toward a description for second language learners. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 88-108). Cambridge: Cambridge University Press.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*(3), 261-77.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, von J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing 24*(1), 37-64.

Elekes, K. (2000). "Please, keep talking": the 'think-aloud' method in second language reading research. *novELTy, 7*(3), 4-14.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data.* Cambridge, MA: MIT Press.

Falus, I. (Ed.) (1996). *Bevezetés a pedagógiai kutatás módszereibe.* Budapest: Keraban Kiadó.

Fekete, H., Major, É., & Nikolov, M. (Eds.) (1999). *English language education in Hungary. A baseline study.* Budapest: British Council.

Fox, J. (2004). Test decisions over time: tracking validity. *Language Testing, 21*(4), 437-465.

Framework curricula for secondary schools. Volume one. (2000). Dinasztia Publishing Company: Budapest. Retrieved on June 18, 2007 from http://www.okm.gov.hu/letolt/nemzet/kerettanterv33.doc

Freidlander, A. (1990). Composing in English: Effects of a first language on writing in English as a second language. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 109-125). Cambridge: Cambridge University Press.

Fullan, M., & Hargreaves, A. (Eds.) (1992). *Teacher development and educational change.* East Sussex, United Kingdom: Falmer Press.

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

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research.* London: Lawrence Erlbaum Associates, Publishers.

Greatorex, J., & Suto, W. M. I. (November, 2005). What goes through a marker's mind? Gaining theoretical insights into the A-level and GCSE marking process. A report of a discussion group at Association for Educational Assessment. Dublin.

Greatorex, J., & Suto, W. M. I. (May, 2006). An empirical exploration of human judgement in the marking of school examinations. Paper presented at the International Association for Educational Assessment Conference, Singapore.

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

Green, T. (2004). Making the grade: score gains on the IELTS writing test. *Research Notes, 16*(4), 9-13.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1992). Holistic writing assessment for LEP students. *Proceedings of Second National Symposium on Limited English Proficient Student.* OBLMA. Retrieved on September 19, 2006 from: http://www.ncela.gwu.edu/pubs/symposia/second/vol2/holistic.htm

Hamps-Lyons, L. (1995). Research on the rating process. Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*(1), 5-16.

Hamps-Lyons, L. (2004). Writing assessment in the world. *Assessing Writing, 9*(1), 1-3.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*(1), 1-9.

Harmer, J. (1991). *The practice of English language teaching.* London: Longman.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of the writing process. In L. W. Gregg, & E.R. Steinberg, *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*(2), 122-159.

Heltai, P. (2001). Communicative language tests, authenticity and the mother tongue. *novELTy, 8*(2), 4-21.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics 25*, pp. 205-227. Retrieved on October 4, 2006 from:

http://journals.cambridge.org/action/displayFulltext?type=6&fid=322811&jid=APL&

Hymes, D. (1972). On communicative competence. In J. Pride, & A. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). New York: Penguin.

Johns, A. M. (1990). L1 composition theories: implications for developing theories of L2 composition. In B. Kroll (Ed.), *Second language writing* (pp. 24-37). Cambridge: Cambridge University Press.

Johns, A. M. (1993). Written argumentation for real audiences: suggestions for teacher research and classroom practice. *TESOL Quarterly, 27*(1), 75-90.

Johnson, S., Linton, P., & Madigan, R. (1994). The role of internal standards in assessment of written discourse. *Discourse Processes, 18*, 231-245.

Jones, N., & Shaw, S. D. (2003). Task difficulty in the assessment of writing: comparing performance across three levels of CELS. *Research Notes, 11*(4), 11-15.

Kitao, S. K., & Kitao, K. (2001). Testing communicative competence. *The Internet TESL Journal.* Retrieved on December 11, 2001 from:

http://iteslj.org/Articles/Kitao-Testing.html.

Krapels, A. R. (1990. An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing* (pp. 37-57). Cambridge: Cambridge University Press.

Krashen, S. D. (1984). *Writing, research, theory and applications.* Oxford: Pergamon Institute of English.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications.* London: Longman.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing* (pp. 140-155). Cambridge: Cambridge University Press.

Kroll, B. (Ed.) (1990). *Second language writing.* Cambridge: Cambridge University Press.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*(1), 4-26.

Leki, I. (2004). Teaching second-language writing: Where we seem to be. In T. Kral (Ed.), *Teacher Development* (pp. 170-178). Washington, DC: English Language Programs Division of the USIA.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud. The Issue of Reactivity in SLA Research Methodology. *Studies in Second Language Acquisition, 26*, 35-57.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: language testing and assessment. *TESOL Quarterly 40*(1), 211-234.

Lumley, T. (2000). *The process of the assessment of writing performance: the rater's perspective.* Unpublished PhD dissertation. Department of Linguistics and Applied Linguistics: The University of Melbourne.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the testers? *Language Testing, 19*(3), 246-276.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Mackey, A., & Gass, S. M. (2005). *Second language research.* Mahwah: Lawrence Erlbaum: Mahwah, New Jersey.

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2000). An approximation to the study of backtracking in L2 writing. *Learning and Instruction, 10*(1), 13-53.

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

McNamara, T. F. (1997). "Interaction" in second language performance assessment: whose performance? *Applied Linguistics, 18* (4), 446-466.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (13-103). New York: Macmillan.

Milanovic, M., & Saville, N. (Eds.) (1996). *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arhem.* Cambridge: Cambridge University Press.

Milanovic. M., Saville, N., & Shuhong, S. (1996). A study of decision-making behaviour of composition markers. In M. Milanovic, & N. Saville (Eds.), *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arhem.* (pp. 92-115). Cambridge: Cambridge University Press.

Nahalka, I. (1996). A statisztikai módszerek pedagógiai alkalmazásának indokai, statisztikai alapfogalmak. In I. Falus, (Ed.), *Bevezetés a pedagógiai kutatás módszereibe* (pp. 343-357). Budapest: Keraban Könyvkiadó.

Nakamura, Y. (May, 2002). A comparison of holistic and analytic scoring methods in the assessment of writing. Paper presented at The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference. Tokyo, Japan: Tokyo Keizai University.

Nemzeti Alaptanterv. (1995). Budapest: Korona.

Nikolov, M. (2006). Why whales have migraine. *Language Learning, 56*(1), 1-51.

Nikolov, M., & Józsa, K. (2003). *Idegen nyelvi készségek fejlettsége angol és német nyelvből a 6. és 10. évfolyamon a 2002/2003-as tanévben. Szakmai beszámoló országos adatok alapján.* Budapest: OKÉV.

Norris, J. M. (2000). Purposeful language assessment. *Forum 38*(1). Retrieved on November 15, 2002 from: http://exchanges.state.gov/forum/vol38/no1/p18.htm.

Nunan, D. (1991). *Language teaching methodology.* New York: Phoenix ELT.

O'Sullivan, B. (2002). Investigating variability in a test of second language writing ability. *Research Notes, 7*(5), 14-17.

Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Education Policy Analysis Archives 8*(16). Retrieved on November 3, 2000 from:

http://epaa.asu.edu/epaa/v8n16/

Penny, J. A. (2003). Reading high stakes writing samples: My life as a reader. *Assessing Writing 8*(3), 192-215.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability. An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143-164.

Perrin, G. (2005). Teachers, testers, and the research enterprise – a slow meeting of minds. *ELT Journal, 59*(2), 144-150.

Pollitt, A. (1991). Giving students a sporting chance. In J.C. Alderson, & B. North (Eds.), *Language testing in the 1990s* (pp. 46-70). London: McMillan.

Pollitt, A., & Murray, N. L. (1995). What raters really pay attention to. In M. Milanovic, & N. Saville, (Eds.), *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arhem.* (pp. 74-92). Cambridge: Cambridge University Press.

Porter, D., & O'Sullivan, B. (1999). The effect of audience age on measured written performance. *System, 27*(1), 65-77.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the teaching of English, 26*(1), 108-122.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19*(2), 229-258.

Részletes vizsgakövetelmény és vizsgaleírás. Angol nyelv (2003). Retreived on June 21, 2007 from: http://www.okm.gov.hu/letolt/kozokt/erettsegitervezet/22/Angol.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research, and Evaluation, 3*(3). Retrieved on July 4, 2003 from:

http://ericae.net/pare/getvn.asp?v=3&n=3

Rudner, L., & Cagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation, 7*(26). Retrieved on July 4, 2003 from:

http://ericae.net/pare/getvn.asp?v=7&n=26.

Sasaki, M. (2000). Toward an empirical model of EFL writing process: An exploratory study. *Journal of Second Language Writing, 9*(3), 259-291.

Scott, T. (2005). Creating the subject of portfolios: reflective writing and the conveyance of institutional prerogatives. *Written Communication, 22*(3), 3-35.

Shaw, S. (2001). Issues in the assessment of second language writing. *Research Notes, 6*(1), 2-6.

Shaw, S. (2002a). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes, 8*(5), 13-17.

Shaw, D. S. (2002b). IELTS writing: revising assessment criteria and scales (Phase 2). *Research Notes, 10*(4), 10-13.

Shaw, S. (2003). Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts. *Research Notes, 11*(3), 7-10.

Shaw, D. S. (2004). IELTS writing: revising assessment criteria and scales (concluding Phase 2). *Research Notes, 15*(3), 9-11.

Shaw, S. (2007). Modelling facets of the assessment of writing within an ESM environment. *Research Notes, 27*(1), 14-19.

Shaw, S., & Jordan, S. (2002). CELS writing: test development and validation activity. *Research Notes, 9*(4), 10-13.

Silva, T. (1990). Second language composition instruction: developments, issues, and directions in ESL. In B. Kroll (Ed.), *Second language writing* (pp. 11-24). Cambridge: Cambridge University Press.

Spaan, M. (1993). The effect of prompt in essay examination. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research. Selected papers from the 1990 Language Testing Research Colloquium.* (pp. 98-123). Alexandria, Virginia: TESOL Publications.

Spencer, S. L, & Fitzgerald, J. (1993). Validity and structure, coherence, and quality measures in writing. *Journal of Reading Behaviour, 25*(2), 209-231.

Spolsky, B. (Ed.) (1999). *Concise encyclopedia of educational linguistics.* Oxford: Elsevier Science ltd.

Stern, L. A., & Solomon, A. (2006). Effective feedback: the road less travelled. *Assessing Writing, 11*(1), 22-41.

Taylor, C. (1996). A study of writing task assigned in academic degree programs: a report on stage 1 of the project. In M. Milanovic, & N. Saville (Eds.), *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arhem* (pp. 115-130). Cambridge: Cambridge University Press.

Taylor, L. (2004). Second language writing assessment: Cambridge ESOL's ongoing research agenda. *Research Notes, 16*(1), 2.

Torrance, H. (1998). Learning from research in assessment: a response to writing assessment – raters' elaboration of the rating task. *Assessing Writing, 5*(1), 31-37.

Tsui, A. B. M. (1996). Learning to teach ESL writing. In D. Freeman, & J. C. Richards *Teacher learning in language teaching* (pp. 97-121). Cambridge: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System, 28*(4), 499-509.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

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

Weir, C. J. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

Weir, C., & Shaw, S. (2006). Defining the constructs underpinning main suite writing tests: a socio-cognitive perspective. *Research Notes, 26*(3), 9-14.

Weissberg, B. (2000). Developmental relationships in the acquisition of English syntax: writng vs. speech. *Learning and instruction, 10*(1), 37-55.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

Wolfe, E. W., Kao, C. W., & Ranney. M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

Wolff, D. (2000). Second language writing: a few remarks on psycholinguistic and instructional issues. *Learning and Instruction, 10*(1), 107-112.

Wong, A. T. Y. (2005). Writers' mental representations of the intended audience and of the rhetorical purpose for writing and the strategies that they employed when they composed. *System, 33*(1), 29-47.

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.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*(1), 79-101.

Appendices

Appendix 6.1
The letter writing task in the pilot study

This text appeared in an Internet magazine for teenagers.

Leslie's Dream Holiday Competition

Imagine you have just come back from a dream holiday. Write us about it.

In your letter to Leslie, the editor of the magazine, write about
- where you travelled and how you got there
- who you went with
- why you went with him or her
- what the place was like
- an interesting thing you did there
- what your next holiday would be like

Write about 150 words.

Dear Leslie,

Appendix 6.2
The rating scale in the pilot study

Grading criteria for assessment of written performance in English and German in the 10th year (translated from the Hungarian version)

	Communicative goal – performance on the 6 content points	**Richness of vocabulary**	**Accuracy and spelling**	**Text organisation**
7-8	The letter is written to a stranger. The candidate writes appropriately about 5 or 6 content points.	Vocabulary use shows wide variety and selection, is appropriate to the task.	There are some grammar and/or spelling mistakes, but the whole text is comprehensible.	The text is well-structured: each issue is dealt with in a separate paragraph; there is logical link between the sentences. Approximately half of the sentences are complex. The script shows letter characteristics well.
5-6	The letter is written to a stranger. The candidate writes appropriately about 4 or 5 content points, or covers about 5 or 6 content points partly.	Vocabulary use is appropriate to the task, shows relatively wide variety and selection.	There are some basic mistakes, but the whole text is comprehensible.	There are no separate paragraphs, but there is some logical link between the sentences, or there are separate paragraphs, but not all sentences are logically linked. There are some complex sentences. The script shows some letter characteristics.
3-4	**3 or 4 points are covered appropriately, or more points are covered partly appropriately.**	**The vocabulary is mostly appropriate to the task and is relevant.**	**Although there are several mistakes, the majority of the text is comprehensible.**	**In most cases there is logical link between the sentences. The text is mostly coherent. Three or four sentence types occur repeatedly, there are complex sentences among them.**
1-2	1 or 2 points covered appropriately, or more points are covered partly appropriately.	Vocabulary is limited and/or irrelevant.	There are a lot of grammar mistakes, only part of the text is comprehensible.	There is a minimal logical link between the sentences. One or two sentence types occur repeatedly. There is no complex sentence.
0	Did not write anything, or wrote some words or 1 or 2 sentences about 1 or 2 points. Handwriting is illegible or the candidate wrote about something else.	Vocabulary is very poor and limited or irrelevant.	The text is not comprehensible because of grammar mistakes and/or spelling mistakes that hinder comprehension.	There is no link between the words and/or sentences. The text is hardly comprehensible.

- It is possible to choose from two scores in each band, except for the 0 band, to arrive at more sophisticated assessment. In case the candidate performance exceeds the given criteria convincingly, the higher score can be awarded. The total score is 32 points; each of the four criteria can be awarded with the maximum of 8 points.
- **The 4 subscores (0-8) should be written on the script to the right margin vertically: they don't have to be added up. If the test paper is left blank, one score can be given: 9.**
- The length of the script counts in deciding how well it achieves the communicative goal. No points should be deducted if the text is longer or shorter as required.
- If the candidate gets a 0 for the achievement of the communicative goal, the script should not have to be assessed, the final score is 0. It can happen in case the candidate wrote a long composition, which is coherent, with good vocabulary and accurately, but failed to fulfil the task achievement: for example, s/he wrote about something else (not about him/herself) or wrote to somebody else. The assessment should begin with comparison of the script to the middle band (in bold), if the criteria are satisfied, advance should be made upwards, if not, down in the table.

Appendix 6.3
The coding scheme for the pilot study

Category	Behaviour	Code
Rating technicalities	Identifies script	**T13**
	Nominates scoring category	**T21**
	Refers to rating technicalities	**T33**
Reading the script	First reads the whole text	**Rd1**
	Rereads the text	**Rd20**
General comments	Refers to comprehension	**G5**
	Remarks on general impression	**G10**
	Remarks on handwriting and legibility	**G11**
	Remarks on layout and length	**G15**
	Expresses personal reaction	**G17**
	Remarks on quality	**G18**
	Remarks on relevance	**G19**
	Refers to candidate's overall proficiency	**G25**
	Concludes rating	**G27**
	Concludes performance	**G29**
	Makes other comment	**G39**
Rating comments	Finalises score	**R9**
	Hesitates	**R12**
	Justifies judgement	**R14**
	Summarises judgement	**R24**
	Repeats score	**R26**
	Identifies other influence	**R28**
	Compares to other performance	**R35**
	Reconsiders evaluation	**R36**
	Forecasts evaluation	**R37**
	Offers solution	**R38**
Communicative goal – fulfilment of the 6 content points	Adds up content points	**CG2**
	Cites from scale	**CG4**
	Evaluates content points	**CG6**
	Reads an example	**CG8**
	Summarises content	**CG23**
	Evaluates communicative goal	**CG31**
	Refers to rubric	**CG34**
	Compares to scale	**CG41**

Category	Behaviour	Code
Richness of vocabulary	Cites from scale	**V4**
	Gives an example	**V8**
	Evaluates vocabulary	**V30**
	Compares to scale	**V41**
Accuracy and spelling	Cites from scale	**Gr4**
	Gives an example	**Gr8**
	Evaluates grammar	**Gr32**
	Compares to scale	**Gr41**
Text organisation	Comments on coherence	**O3**
	Cites from scale	**O4**
	Comments on letter conventions	**O7**
	Gives an example	**O8**
	Comments on paragraphing	**O16**
	Comments on sentence variety	**O22**
	Evaluates organisation	**O40**
	Compares to scale	**O41**

Appendix 6.4
Number of utterances made during the rating process in the pilot study

Category	Behaviour	Code	Number of utterances
Scoring technicalities	Identifies script	**T13**	**74**
	Nominates scoring category	**T21**	**156**
	Refers to rating technicalities	**T33**	**18**
Reading the script	First reads the whole text	**Rd1**	**36**
	Rereads the text	**Rd20**	**11**
General comments	Refers to comprehension	**G5**	**38**
	Remarks on general impression	**G10**	**19**
	Remarks on handwriting and legibility	**G11**	**17**
	Remarks on layout and length	**G15**	**37**
	Expresses personal reaction	**G17**	**44**
	Remarks on quality	**G18**	**43**
	Remarks on relevance	**G19**	**37**
	Refers to candidate's overall proficiency	**G25**	**4**
	Concludes rating	**G27**	**5**
	Concludes performance	**G29**	**14**
	Makes other comment	**G39**	**34**
Rater behaviour	Finalises score	**R9**	**254**
	Hesitates	**R12**	**31**
	Justifies judgement	**R14**	**47**
	Summarises judgement	**R24**	**37**
	Repeats score	**R26**	**28**
	Identifies other influence	**R28**	**7**
	Compares to other performance	**R35**	**23**
	Reconsiders evaluation	**R36**	**14**
	Forecasts evaluation	**R37**	**6**
	Offers solution	**R38**	**4**
Communicative goal	Adds up content points	**CG2**	**25**
	Cites from scale	**CG4**	**6**
	Evaluates content points	**CG6**	**70**
	Reads an example	**CG8**	**31**
	Summarises content	**CG23**	**42**
	Evaluates communicative goal	**CG31**	**22**
	Refers to rubric	**CG34**	**10**
	Compares to scale	**CG41**	**7**

Category	Behaviour	Code	Number of utterances
Richness of vocabulary	Cites from scale	**V4**	**28**
	Gives an example	**V8**	**20**
	Evaluates vocabulary	**V30**	**34**
	Compares to scale	**V41**	**20**
Accuracy and spelling	Cites from scale	**Gr4**	**12**
	Gives an example	**Gr8**	**1**
	Evaluates grammar	**Gr32**	**65**
	Compares to scale	**Gr41**	**9**
Text organisation	Comments on coherence	**O3**	**44**
	Cites from scale	**O4**	**6**
	Comments on letter conventions	**O7**	**27**
	Gives an example	**O8**	**7**
	Comments on paragraphing	**O16**	**37**
	Comments on sentence variety	**O22**	**48**
	Evaluates organisation	**O40**	**15**
	Compares to scale	**O41**	**5**
Total			**1,629**

Appendix 6.5
A sample from EngR1 transcript in the pilot study (translated from Hungarian)

1	Number of script: 048015202.	T13
2	First, I'll read the letter.	Rd1
3	I can see that s/he wrote on the first page only.	G15
4	And quite a lot.	G15
5	There is greeting and signature as well.	O7
6	It seems that s/he has written almost about everything.	G10
7	Now I'll check how many points s/he has covered.	T33
8	S/he wrote about the place he went, and who s/he travelled with and why. S/he also wrote about what the place looked like, what was interesting, and what s/he was doing there. In addition, s/he wrote where s/he would like to go next.	CG6
9	Now, looking at the scale, I can see that s/he covered six content points.	CG41
10	Now, I am going to reread the letter and check the points again.	Rd20
11	I'll give 8 points for that.	R9
12	I am looking at vocabulary.	T21
13	If I start looking at the middle column, it is: "mostly appropriate to the task and relevant"	V4
14	It is more than that.	V30
15	Now, I am looking how much higher I can go.	V41
16	I think it is: "wide variety and selection, is appropriate to the task".	V4
17	So, I'll give 8 points for that.	R9
18	Accuracy and spelling.	T21
19	There are some mistakes there.	G18
20	I am rereading the letter again.	Rd20
21	I think "there are some grammar mistakes but the whole text is comprehensible".	Gr4
22	So, s/he will get 8 points for that as well.	R9
23	As far as text organization is concerned.	T21
24	I am going to look at the middle section of the scale or even higher.	O41
25	There are no paragraphs, yes, there are no paragraphs.	O16
26	There is some logic between the sentences.	O3
27	So, I'll give 6 points for that.	R9
28	Because there are no paragraphs, but there are logical links in it.	R14
29	So, s/he will get 6 points.	R26
30	Number of the next script: 048016202.	T13
31	First, I'll check how much s/he wrote.	G15
32	Uhh	G17

Appendix 7.1

The writing task in the main study

You spent your holiday in Britain and stayed with a British friend. You arrived back home last week and now you are writing a letter to your friend.

In your letter

- thank your friend for the time you had together
- write about journey home
- write about what presents you brought and for whom
- invite your friend for the next holiday
- write about programmes you are planning for his/her stay in Hungary

Write a letter of about 150 words to your friend.

Dear Pat,

Appendix 7.2
The rating scale in the main study

POINTS	TASK ACHIEVEMENT	VOCABULARY	GRAMMAR	ORGANIZATION
6	☐ Achieves communicative goal; the letter is for a friend ☐ All 5 content points covered	☐ Wide range of appropriate words and expressions	☐ Only one or two inaccuracies occur ☐ Structures correspond to task	☐ The layout fully corresponds the task ☐ There is clear logical link between all text levels
5				
4	☐ Communicative goal mostly achieved ☐ Almost all content points covered	☐ Good and appropriate range of words and expressions	☐ There are some inaccuracies but the whole text is comprehensible ☐ Some variety in structures	☐ The layout reminds of a letter ☐ There is some link at most text levels
3				
2	☐ Strain on the reader to comprehend ☐ 2 or 3 content points covered	☐ Basic words and expressions	☐ Basic mistakes hinder comprehension ☐ No variety in structures	☐ The layout is inappropriate ☐ There is no clear logical link between the elements of the text
1	☐ Communicate goal is difficult to comprehend ☐ 1 content point covered	☐ Very limited range of words	☐ Only part of text is comprehensible ☐ Most structures inaccurate	☐ The layout is messy ☐ Logic is missing between the different elements of the text
0	☐ Did not write anything or just some words ☐ Misunderstood the task			

Appendix 7.3
The letter to the students in the main study

Dear Student,
Thank you very much for helping me in my research. As explained in the seminar, you are kindly asked to produce a think-aloud protocol. The following checklist will help you in case of a problem.
Please, read it carefully before starting rating and refer to it any time you need

☐ Before starting the assessment/recording, please make sure:
- you have the ten scripts: **Scripts 1, 2. 3, 4, 5, 6, 7, 8, 9, 10**
- the audio-recording is of good quality and satisfactory for making a transcript
- you have studied the task carefully: what the students had to write and how much
- you check the content points and the style (informal letter)
- you are familiar with the scale and can use it with ease
- your name is on the disc, audiocassette and the transcript as well

☐ CHECK YOUR SEMINAR NOTES AND THINK OF ANY PROBLEMS THAT YOU MAY FACE (you can contact me: bkata@sol.cc.u-szeged.hu)

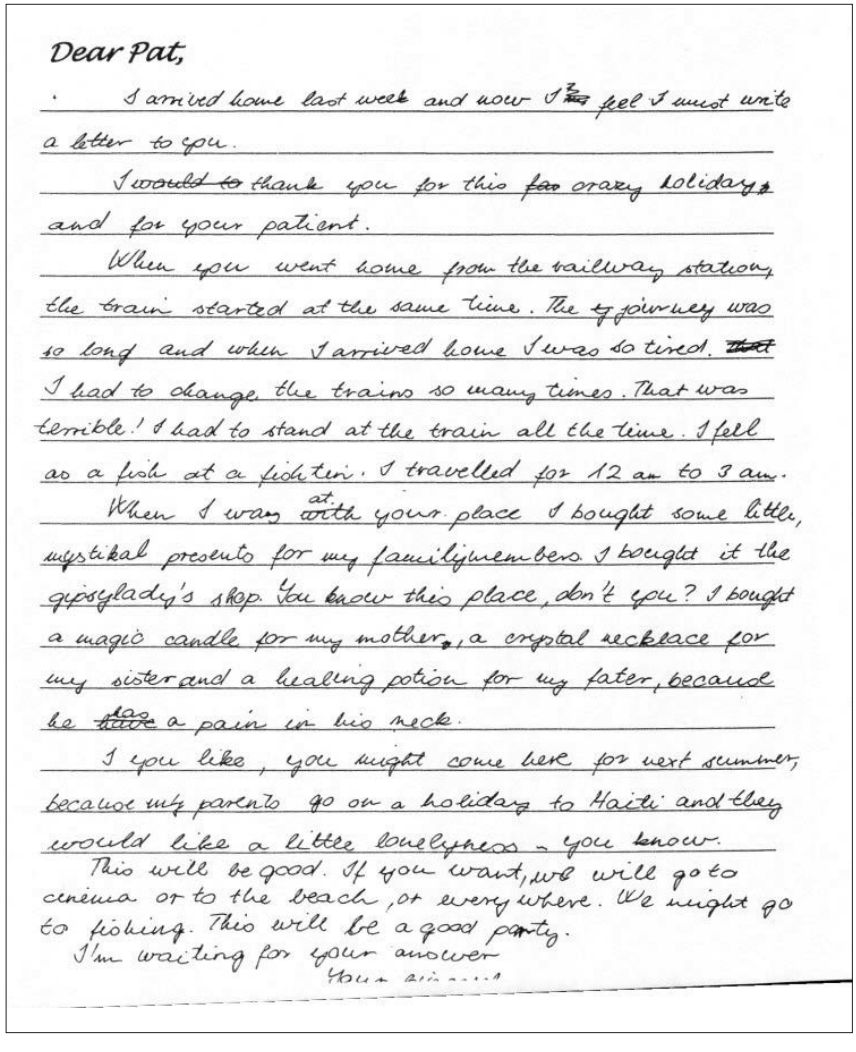
☐ During assessment/recording
- You are supposed to read each script twice: during first reading assess task achievement only; then read it again to assess the other aspects.
- Don't forget to identify the script: each has a number in the top right corner: e.g.: ScriptN1, etc.
- Keep talking and vocalize each thought you have (as we agreed, the language doesn't count).
- Fill in the score sheet as you are doing the assessment.

☐ After the assessment/recording
- produce the transcript on the floppy disc and a hard copy is also needed
- fill in the feedback questionnaire

EACH ENVELOPE CONTAINS:
an audiocassette
a floppy disc
a rating scale
a score grid/sheet
10 scripts
a feedback sheet

Appendix 7.4
The ten scripts in the main study

Script N1



Dear Pat,

I arrived home last week and now I feel I must write a letter to you.

I would to thank you for this crazy holiday and for your patient.

When you went home from the railway station, the train started at the same time. The journey was so long and when I arrived home I was so tired. I had to change the trains so many times. That was terrible! I had to stand at the train all the time. I felt as a fish at a fishtin. I travelled for 12 am to 3 am.

When I was at with your place I bought some little, mystikal presents for my familijmembers. I bought it the gipsylady's shop. You know this place, don't you? I bought a magic candle for my mother, a crystal necklace for my sister and a healing potion for my fater, because he has a pain in his neck.

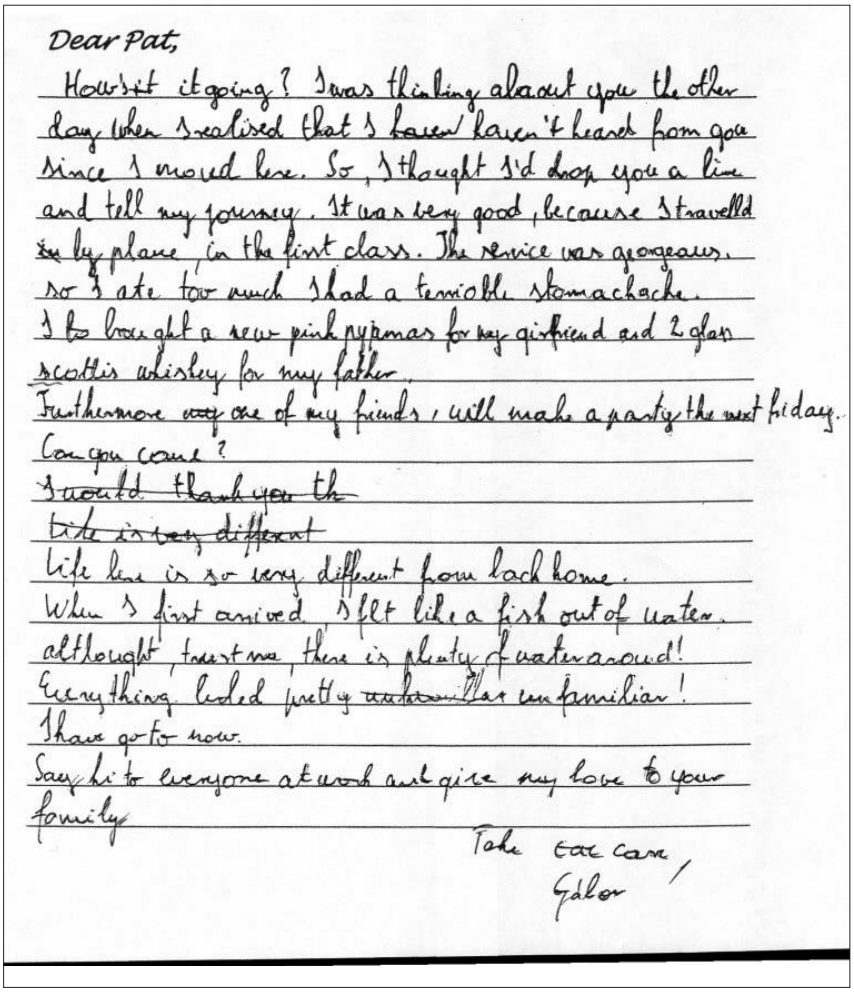
I you like, you might come here for next summer, because my parents go on a holiday to Haiti and they would like a little lonelyness - you know.

This will be good. If you want, we will go to cinema or to the beach, or every where. We might go to fishing. This will be a good party.

I'm waiting for your answer

Your

Script N2



Dear Pat,

How's it going? I was thinking about you the other day when I realised that I haven't heard from you since I moved here. So, I thought I'd drop you a line and tell my journey. It was very good, because I travelld ~~in~~ by plane, in the first class. The service was gorgeous. so I ate too much I had a terrible stomachache. I to brought a new pink pyjamas for my girlfriend and 2 glas scottis whisley for my father.

Furthermore ~~my~~ one of my friends, will make a party the next friday. Can you come?

~~I would thank you th~~

~~life is very different~~

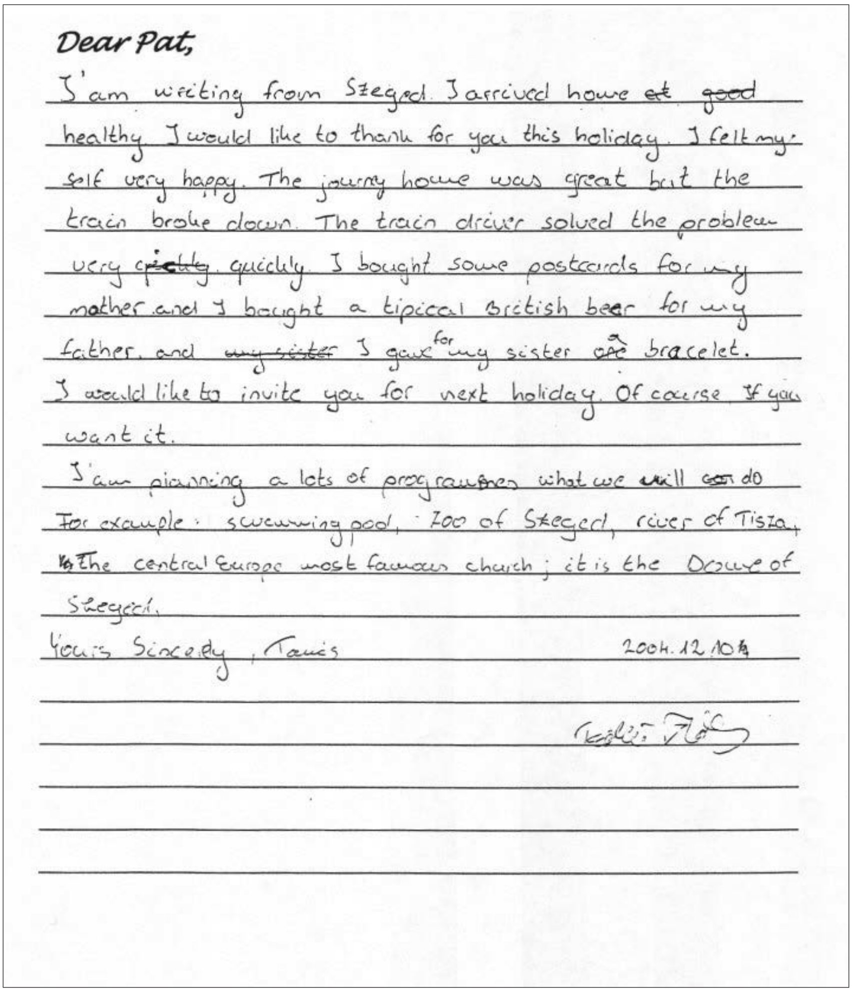
Life here is so very different from back home. When I first arrived I felt like a fish out of water. although, trust me, there is plenty of water around! Everything looked pretty ~~unfamiliar~~ unfamiliar! I have go for now.

Say hi to everyone at work and give my love to your family

Take Eat care

Gabor

Script N3



Dear Pat,

I'am writing from Szeged. I arrived home ~~et~~ ~~good~~ healthy. I would like to thank for you this holiday. I felt my-self very happy. The journy home was great but the train broke down. The train driver solved the problem very ~~qickly~~ quickly. I bought some postcards for my mother and I bought a tipical British beer for my father, and ~~my sister~~ I gave for my sister ~~a~~ a bracelet. I would like to invite you for next holiday. Of course if you want it.

I'am planning a lots of programmes what we will ~~can~~ do. For example: swimming pool, Zoo of Szeged, river of Tisza, The central Europe most famous church; it is the Dome of Szeged.

Yours Sincerely, Tamás 2004.12.10.

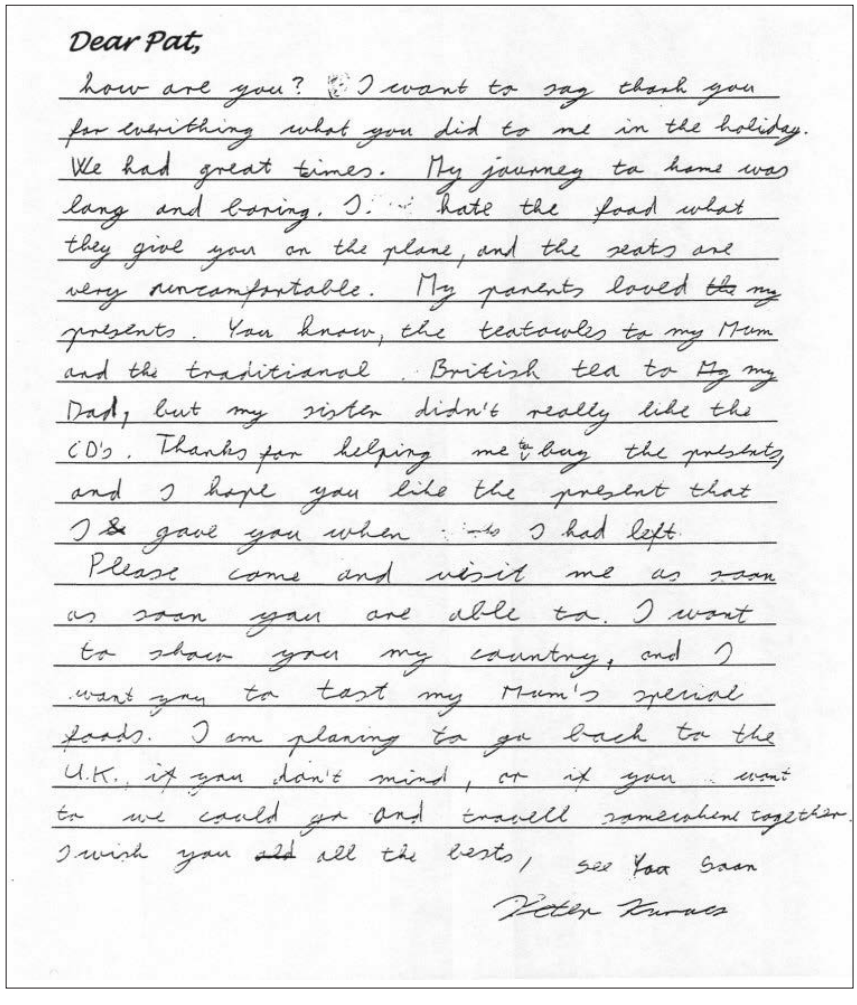
Script N4

Dear Pat,

I thank for the nice time. I felt very good on the last week in Britain. ~~I hope~~ I hope that you ~~feel~~ felt also good. ~~Wh~~ We were ~~very~~ many places that ~~was~~ very beautiful because I like very much the sightseeing. ~~when I must go home~~ when I had to go home, the flight was very interesting, and spectacular because I saw many countries, and lands. But ~~it was the flight was very we~~ I spent many time on the board of the plane. when ~~I~~ I arrived home, ~~it was a good~~ I told my ~~family~~ talked about my history to the family. It was a very beautiful feeling. I bought ~~very much present~~ lots ~~lots~~ of presents. I bought a chockolate to my brother, ~~brittish~~ british tea to my father and ~~a~~ a shoes ~~for~~ to my mother. Would you like to come next holiday ~~to me~~ me? ~~I think~~ I thought about it. I planned lots of programmes, such as going to cinema, ~~etc~~. We will soon go to the ~~park~~ park, go ~~dancing~~ dancing and etc. I hope we will meet soon.

Best regards,
Kovács Gábor

Script N5



Dear Pat,

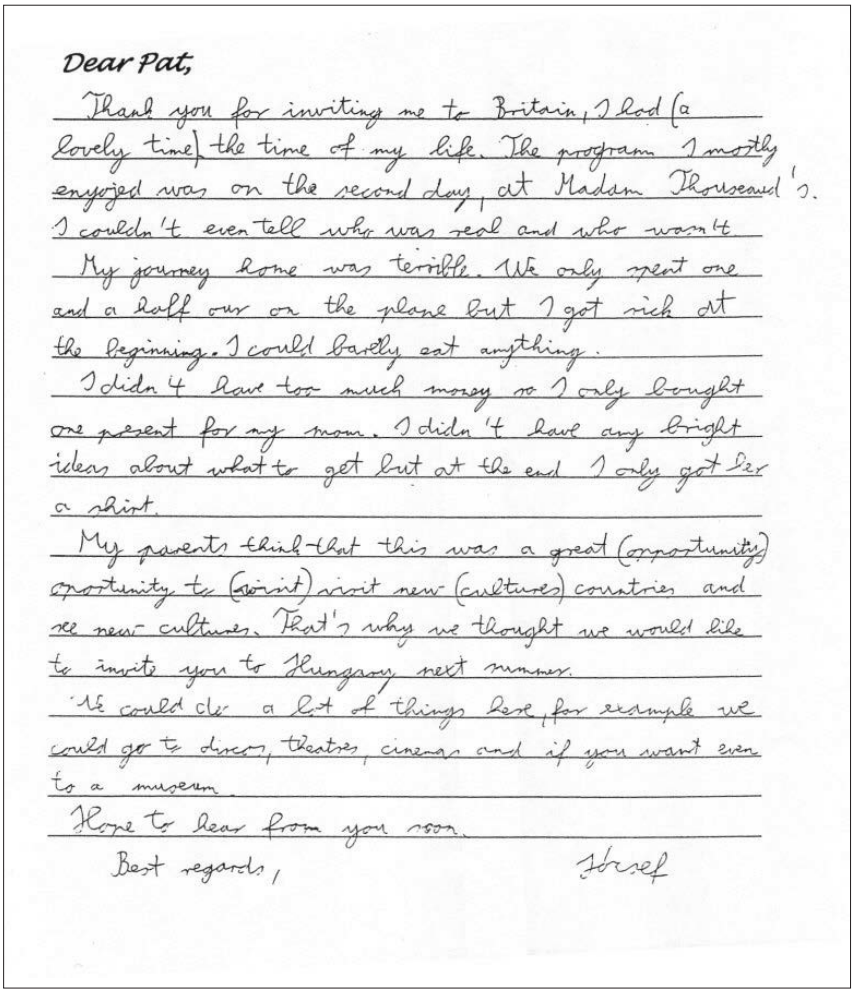
how are you? I want to say thank you for everithing what you did to me in the holiday. We had great times. My journey ta home was lang and boring. I ~~hate~~ the food what they give you on the plane, and the seats are very uncomfortable. My parents loved ~~the~~ my presents. You know, the teatowles to my Mum and the traditianal British tea to ~~My~~ my Dad, but my sister didn't really like the CD's. Thanks for helping me ~~to~~ buy the presents and I hope you like the present that I ~~&~~ gave you when I had left.

Please come and visit me as soon as soon you are able ta. I want to show you my country, and I want you to tast my Mum's special foods. I am planing ta go back to the U.K. if you don't mind, or if you want ta we could go and travell somewhere together.

I wish you ~~all~~ all the bests, see Yoa Soon

Peter Kovacs

Script N6



Dear Pat,

Thank you for inviting me to Britain, I had (a lovely time) the time of my life. The program I mostly enjoyed was on the second day, at Madam Thousand's. I couldn't even tell who was real and who wasn't.

My journey home was terrible. We only spent one and a half our on the plane but I got sick at the beginning. I could hardly eat anything.

I didn't have too much money so I only bought one present for my mom. I didn't have any bright ideas about what to get but at the end I only got her a shirt.

My parents think that this was a great (opportunity) opportunity to (visit) visit new (cultures) countries and see new cultures. That's why we thought we would like to invite you to Hungary next summer.

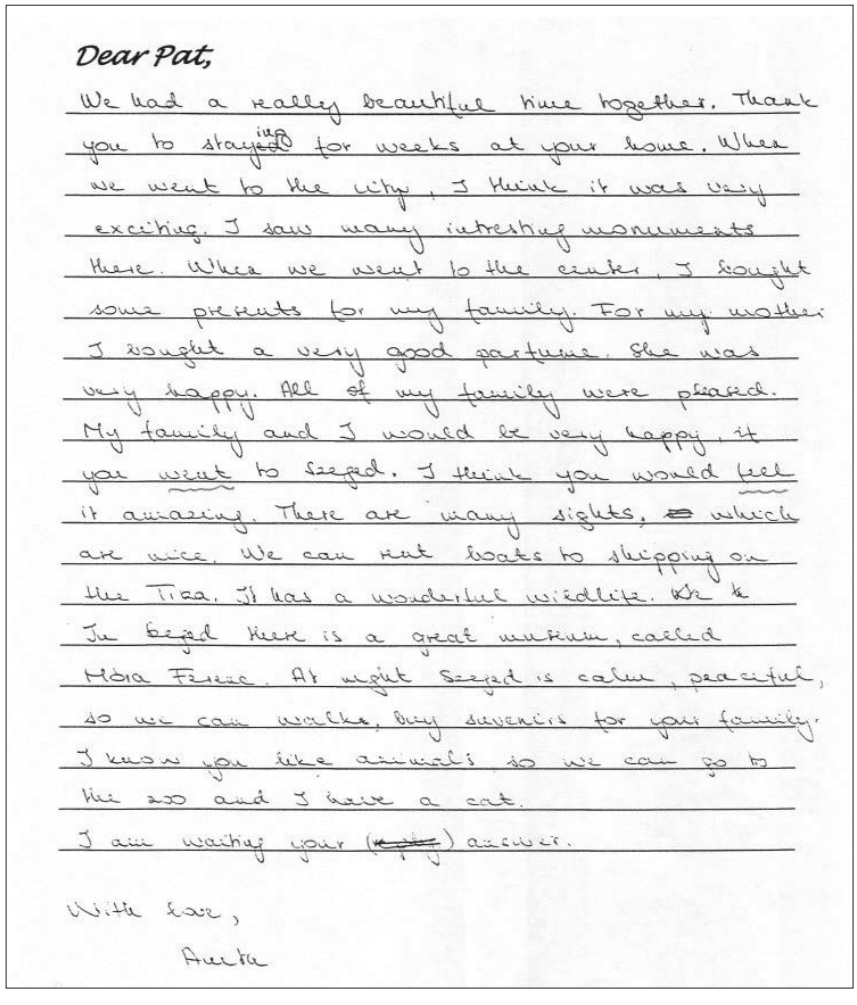
We could do a lot of things here, for example we could go to discos, theatres, cinemas and if you want even to a museum.

Hope to hear from you soon.

Best regards,

Yourself

Script N7



Dear Pat,

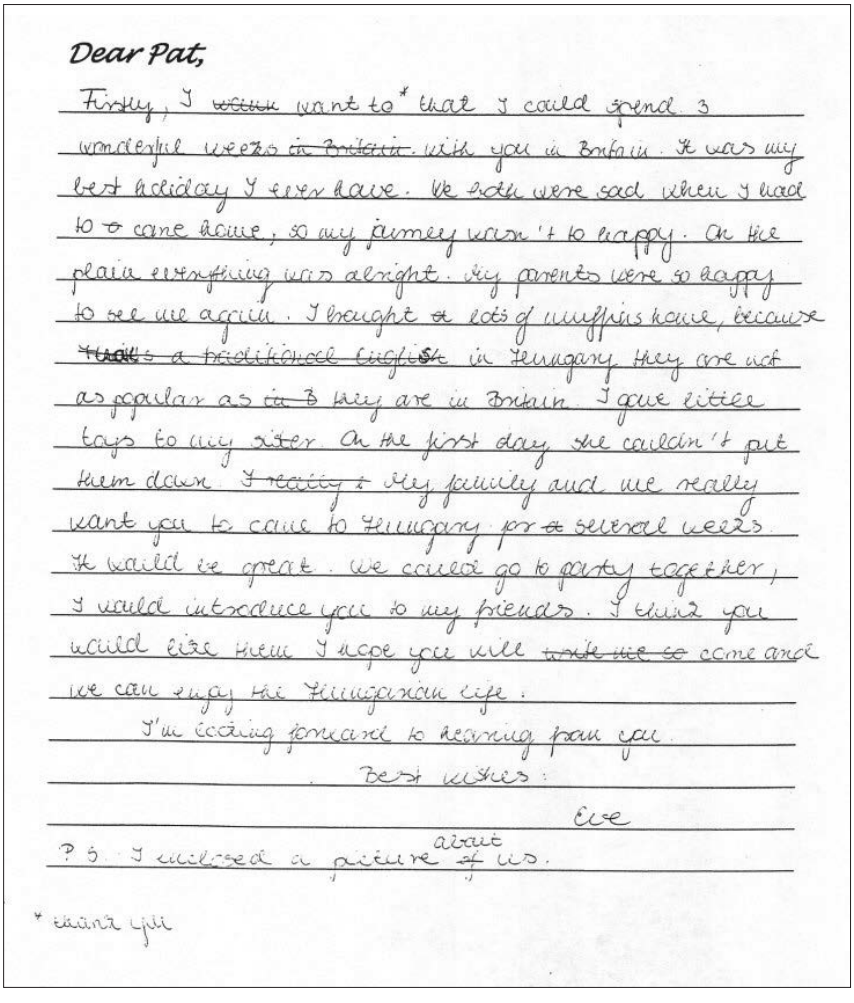
We had a really beautiful time together. Thank you to stay~~ed~~ for weeks at your home. When we went to the city, I think it was very exciting. I saw many intresting monuments there. When we went to the center, I bought some presents for my family. For my mother I bought a very good perfume. She was very happy. All of my family were pleased. My family and I would be very happy, if you went to Szeged. I think you would feel it amazing. There are many sights, ~~=~~ which are nice. We can rent boats to shipping on the Tisa. It has a wonderful wildlife. ~~It~~ In Szeged there is a great museum, called Móra Ferenc. At night Szeged is calm, peaciful, so we can walks, buy souvenirs for your family. I know you like animals so we can go to the zoo and I have a cat.

I am waiting your (~~answer~~) answer.

With love,

Anita

Script N8



Dear Pat,

Firstly, I ~~wanna~~ want to* that I could spend 3 wonderful weeks ~~in Britain~~ with you in Britain. It was my best holiday I ever have. We both were sad when I had to ~~o~~ come home, so my journey wasn't to happy. On the plain everything was alright. My parents were so happy to see me again. I brought ~~a~~ lots of muffins home, because ~~muffin~~ ~~s a traditional English~~ in Hungary they are not as popular as ~~in~~ B they are in Britain. I gave little toys to my sister. On the first day she couldn't put them down. ~~I really +~~ My family and me really want you to come to Hungary for ~~a~~ several weeks. It would be great. We could go to party together, I would introduce you to my friends. I think you would like them. I hope you will ~~write me so~~ come and we can enjoy the Hungarian life.

I'm looking forward to hearing from you.

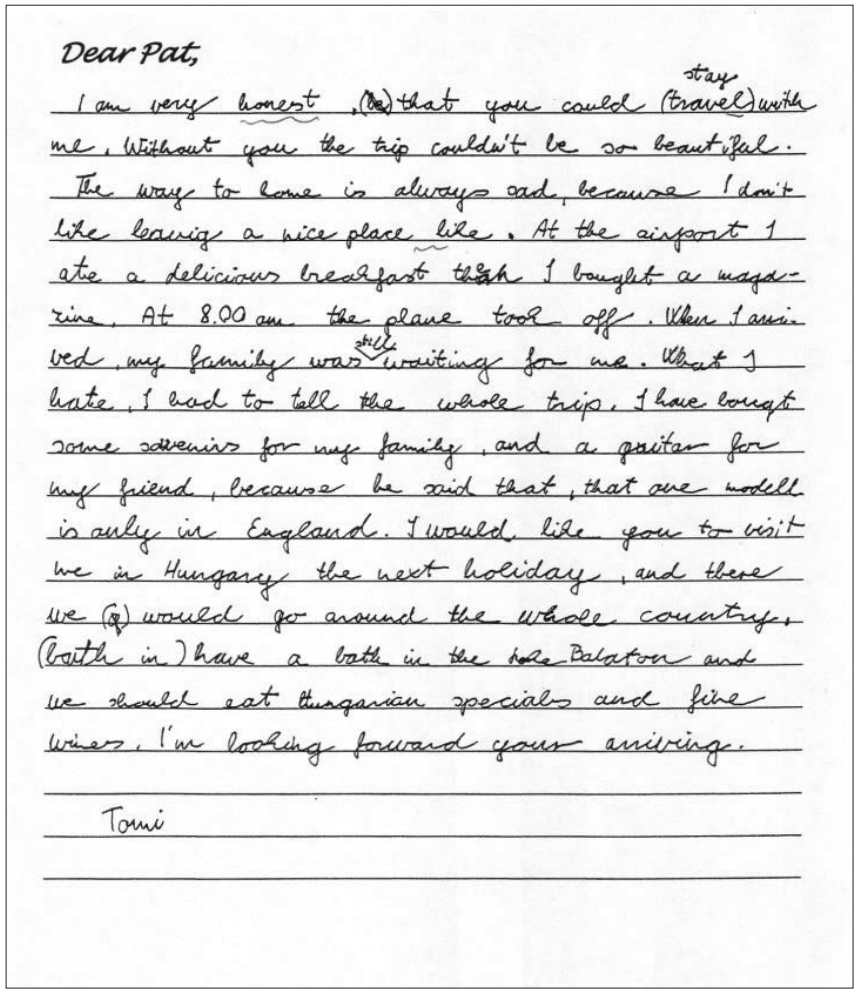
Best wishes:

Eve

P.S. I enclosed a picture about us.

* want till

Script N9

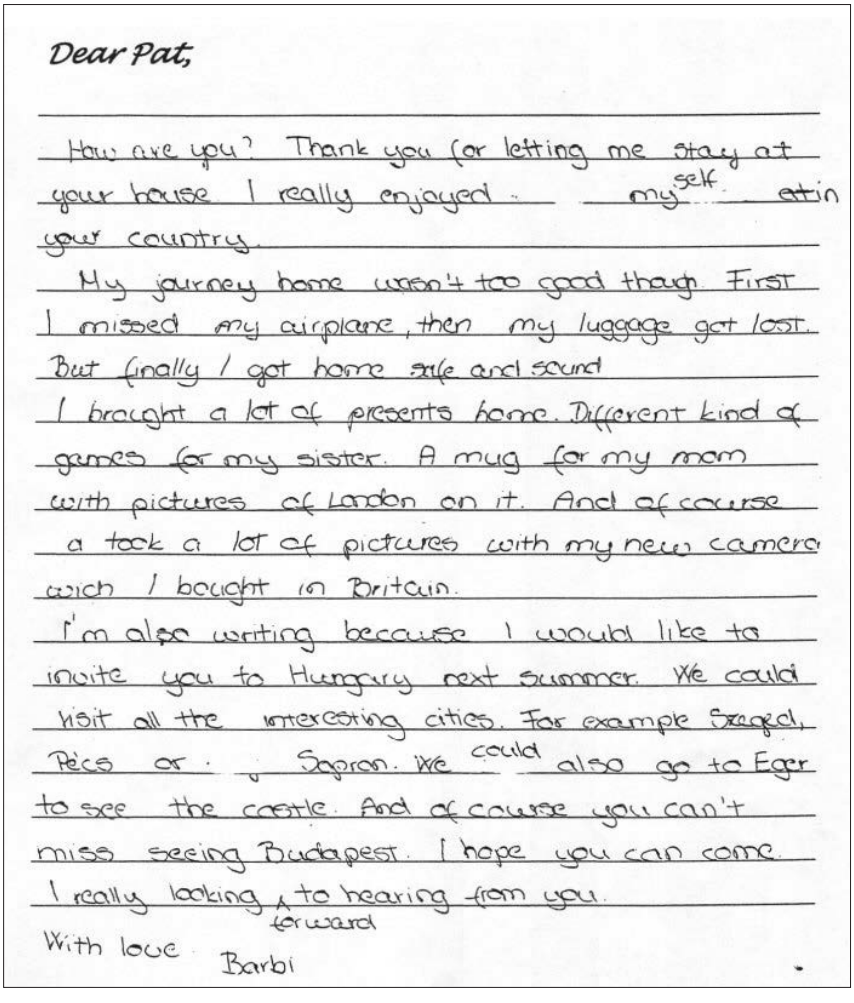


Dear Pat,

I am very honest ,(be) that you could (travel) stay with me. Without you the trip couldn't be so beautiful. The way to home is always sad, because I don't like leaving a nice place like. At the airport I ate a delicious breakfast then I bought a magazine. At 8.00 am the plane took off. When I arrived, my family was still waiting for me. What I hate, I had to tell the whole trip. I have bought some souvenirs for my family, and a guitar for my friend, because he said that, that one modell is only in England. I would like you to visit me in Hungary the next holiday, and there we (a) would go around the whole country, (bath in) have a bath in the lake Balaton and we should eat hungarian specials and fine wines. I'm looking forward your arriving.

Tomi

Script N10



Dear Pat,

How are you? Thank you for letting me stay at your house I really enjoyed myself in your country.

My journey home wasn't too good though. First I missed my airplane, then my luggage got lost. But finally I got home safe and sound

I brought a lot of presents home. Different kind of games for my sister. A mug for my mom with pictures of London on it. And of course a took a lot of pictures with my new camera wich I bought in Britain.

I'm also writing because I would like to invite you to Hungary next summer. We could visit all the interesting cities. For example Szeged, Pécs or Sopron. We could also go to Eger to see the castle. And of course you can't miss seeing Budapest. I hope you can come. I really looking forward to hearing from you.

With love Barbi

Appendix 7.5
The score sheet in the main study

SCORE SHEET RATER'S NAME:.....

SCRIPT NUMBER	TASK ACHIEVEMENT	VOCABULARY	GRAMMAR	ORGANISATION	TOTAL
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Appendix 7.6
The feedback sheet in the main study

Make comments on the

(a) course you had on language testing in general
(b) the training you had for the rating of written performance task
(c) the rating task itself (How do you feel after completion?)

Thank you very much for cooperation, Bukta Katalin

Appendix 7.7
The course description for the elective seminar course in Testing in ELT

Testing in ELT. Methodology Option

Instructor: Bukta Katalin
Time: Monday 18-20
Venue: 107

The course focuses on testing in ELT classes and aims at familiarizing teacher trainees with the basic theories in language testing. In addition, test construction and evaluation techniques will be presented. Bearing in mind communicative language teaching, testing is based on the four skills: reading, writing, speaking and listening. Testing vocabulary and grammar will also be included. The current state of art in language testing in Hungary will also be discussed.

Course outline

1. Introduction. Awareness-raising: Why is testing important?
2. Teaching and testing in the language classroom
3. Basic issues in test characteristics
4. Principles of test design
5. Testing vocabulary
6. Testing grammar
7. Testing reading comprehension
8. Testing speaking
9. Testing listening comprehension
10. Testing writing
11. Rater training for testing writing
12. Rater training for testing writing
13. Think-aloud procedure
14. Rater training for testing writing

Basic literature

Alderson, C. J., C. Clapham and D. Wall (1995): *Language Test Construction and Evaluation*. Cambridge: CUP.
Bachman, L. F. (1990): *Fundamental Considerations in Language Testing.* Oxford: OUP.
Heaton, J. B. (1988): *Writing English Language Tests. London:* Longman.

Assessment

☐ Active participation in the seminars

☐ Assignment as described in the seminar

Appendix 7.8
A sample verbal protocol transcript in the main study

R1 Script N1 231 words

Task achievement. All 5 content points are covered in this composition, but the end is short, the two last content points are short. But the communicative goal is mostly achieved. So I would give 5 points for task achievement.
Vocabulary is good. There are good words and expressions, but the expressions are not always appropriate, for example „I fell as a fish at a fishtin" is not correct, but there are some words which don't belong to the basic vocabulary, for example „mystical presents", „crystal necklace", „healing potion", „magic candle". And the word patient is not correctly used in the sentence „I thank you for this crazy holiday and for your patient." So for vocabulary I would give 4 points.
Grammar. There are some mistakes, but the whole composition is comprehensible. Some mistakes are for example „I fell as a fish at a fishtin."or „I travelled for 12 am to 3 am." is not correct. „I bought it the gipsylady's shop." There is a spelling mistake in this word „gipsylady". There is a spelling mistake in „fater". „We might go to fishing" is also incorrect. But there is a variety in structures, so I would give 4 points for grammar.
Organization. The layout is good. There are many paragraphs, it is well organized. There are some linking elements, for example *and, because, or*. So I would give 5 points for organization.

Appendix 7.9
The coding scheme in the main study

Behaviour type		
1 Identifies a script/criterion		
A Rating focus	Rating aspect	Code Comment
	A1 Task achievement	
		A1a Compares text to scale descriptor 1 communicative goal
		A1b Compares text to scale descriptor 2 content points
		A1c Evaluates aspect in own words
		A1d Adds own criterion
		A1e Chooses score
		A1f Adds reason why that score
		A1g Revises decision
		A1h Identifies error
		A1i Refers to lack of detail
		A1j Changes focus/switches to different criterion
		A1k Finalises the score
	A2 Vocabulary	
		A2a Compares text to scale descriptor 1: range
		A2b Compares text to scale descriptor 2: appropriacy
		A2c Evaluates aspect in own words
		A2d Adds own criterion
		A2e Chooses score
		A2f Adds reason why that score
		A2g Revises decision
		A2h Identifies error
		A2i Refers to lack of detail
		A2j Changes focus/switches to different criterion
		A2k Finalises the score
	A3 Grammar	
		A3a Compares text to scale descriptor 1: accuracy
		A3b Compares text to scale descriptor 2: structures
		A3c Evaluates aspect in own words
		A3d Adds own criterion
		A3e Chooses score
		A3f Adds reason why that score
		A3g Revises decision

		A3h Identifies error
		A3i Refers to lack of detail
		A3j Changes focus/switches to different criterion
		A3k Finalises the score
	A4 Organisation	
		A4a Compares text to scale descriptor 1 layout
		A4b Compares text to scale descriptor 2 links
		A4c Evaluates the aspect in own words
		A4d Adds own criterion
		A4e Chooses score
		A4f Adds reason why that score
		A4g Revises decision
		A4h Identifies error
		A4i Refers to lack of detail
		A4j Changes focus/switches to different criterion
		A4k Finalises the score
B Reading focus	Rating aspect	Code Reading target
	B1 Task achievement	
		B1a Scale
		B1b Script: more words
		B1c Summarises script
		B1d Example: one word
		B1e Rubric
	B2 Vocabulary	
		B2a Scale
		B2b Script: more words
		B2c Summarises script
		B2d Example: one word
		B2e Rubric
	B3 Grammar	
		B3a Scale
		B3b Script: more words
		B3c Summarises script
		B3d Example: one word
		B3e Rubric
	B4 Organisation	
		B4a Scale
		B4b Script: more words
		B4c Summarises script
		B4d Example: one word
		B4e Rubric

C Other comments (own focus)		Code Comment type
		Ca Reflects on length
		Cb Reflects on eligibility, tidiness
		Cc Reflects on quality of script
		Cd Comments on overall impression
		Ce Comments on comprehensibility
		Cf Comments on observation of student proficiency
		Cg Corrects error
		Ch Reflects own feeling
		Ci Reflects on relevance of content
		Cj Meditates on student intention
		Ck Suggests solution
		Cl Compares to other script/student/score
		Cm Expresses uncertainty
D Remarks on the rating process		Code Comment type
		Da Repeats score
		Db Adds up scores
		Dc Unclear reasoning
		Dd Comments on rating process

Appendix 7.10
A sample of a coded protocol in the main study

R9 Script N2

7	Now let's see the next one which is script number 2	1
8	And after the first reading,	Dd
9	it seems that the content points here also... all four... **all five content points were covered**	B1a
10	**and the communicative goal was also achieved**	B1a
11	in this case, so for task achievement	1
12	it is 6.	A1e
13	As for the vocabulary items used in this composition,	1
14	there were very nice phrases and expressions,	A2a
15	and very good use of colloquial language... that	A2b
16	I found examples: *How is it going?*	B2b
17	a typical colloquial question.	Ch
18	Plus phrases like *drop you a line*	B2b
19	*gorgeous*	B2d
20	*fish out of water*	B2b
21	which is a very happy remark....	Ch
22	Thus I gave 6 points for vocabulary.	A2e
23	And as for the grammar,	1
24	**there are inaccuracies** but that does not cause problems in understanding.	B3a
25	Though several spelling mistakes occur	A3c
26	like *scottis*	B3d
27	and *georgeous* were spelled inaccurately.	B3d
28	And there were also grammatical inaccuracies in the text.	A3a
29	For example, *tell my journey*	B3b
30	instead of tell about something.	Cg
31	*And I have go to now*	B3b
32	umm I think he wanted to use the 'have to' structure' (I have to go now)	Ck
33	but I don't think it was an intentional mistake, or it wasn't an error. It was just a mistake.	Ch
34	So I gave 4 points for grammar	A3e
35	due to these mistakes.	A3f
36	As for the organization,	1
37	I gave three points,	A4e
38	for **the layout is not properly organized**, and the paragraphing is very awkward...	B4a
39	**but link words are used...**	B4a
40	hmm so that makes the organization points 3	A4e
41	which is altogether 19 points for script number 2.	Db

Appendix 7.11
Competent raters' total scores and rankings of the ten scripts in parentheses in the main study

Rater	N1(rank)	N2(rank)	N3(rank)	N4(rank)	N5(rank)	N6(rank)	N7(rank)	N8(rank)	N9(rank)	N10(rank)	Range
R1	18(1)	11(5)	12(4)	10(6)	15(3)	18(1)	12(4)	17(2)	12(4)	17(2)	18-10(1-6)
R4	12(5)	13(4)	13(4)	9(7)	12(5)	22(1)	11(6)	16(3)	13(4)	20(2)	22-9(1-7)
R5	17(3)	0(9)	14(4)	9(8)	12(6)	19(1)	13(5)	13(5)	11(7)	18(2)	19-0(1-9)
R7	18(2)	15(4)	14(5)	10(8)	11(7)	13(6)	16(3)	20(1)	11(7)	15(4)	20-10(1-8)
R9	18(3)	19(2)	15(6)	12(7)	16(5)	24(1)	16(5)	17(4)	12(7)	19(2)	24-12(1-7)
R10	11(6)	11(6)	18(4)	11(6)	15(5)	22(2)	21(3)	21(3)	9(7)	23(1)	23-9(1-7)
R11	12(7)	13(6)	14(5)	12(7)	19(3)	19(3)	16(4)	22(1)	20(2)	19(3)	22-12(1-7)
R12	14(4)	14(4)	10(6)	10(6)	13(5)	22(2)	14(4)	16(3)	10(6)	23(1)	23-10(1-6)
R13	18(2)	10(7)	12(5)	12(5)	13(4)	19(1)	13(4)	15(3)	11(6)	18(2)	19-10(1-7)
R15	20(3)	6(8)	23(1)	12(7)	15(6)	21(2)	19(4)	18(5)	15(6)	20(3)	23-6(1-8)
R16	19(4)	10(8)	17(5)	10(8)	22(2)	21(3)	16(6)	22(2)	14(7)	23(1)	23-10(1-8)
R17	15(5)	15(5)	10(7)	8(8)	10(7)	24(1)	23(2)	18(4)	11(6)	19(3)	24-8(1-8)
R18	19(5)	20(4)	16(6)	11(8)	16(6)	23(1)	21(3)	19(5)	14(7)	22(2)	23-11(1-8)
R19	17(5)	9(7)	17(5)	19(4)	24(1)	22(2)	20(3)	17(5)	10(6)	24(1)	24-9(1-7)
RR2	21(2)	15(6)	13(7)	18(4)	16(5)	24(1)	16(5)	19(3)	15(6)	21(2)	24-13(1-7)
RR5	19(3)	12(6)	16(4)	15(5)	19(3)	23(1)	19(3)	22(2)	19(3)	22(2)	23-12(1-6)
RR6	18(3)	15(5)	17(4)	12(6)	17(4)	20(2)	17(4)	17(4)	18(3)	24(1)	14-12(1-6)
RR7	18(4)	13(7)	11(9)	15(6)	17(5)	20(2)	19(3)	22(1)	12(8)	19(3)	22-11(1-9)
RR9	21(3)	19(4)	13(8)	18(5)	18(5)	24(1)	16(7)	19(4)	17(6)	23(2)	24-13(1-8)
RR10	20(3)	18(5)	17(6)	16(7)	17(6)	21(2)	22(1)	20(3)	19(4)	22(1)	22-16(1-7)
RR14	17(7)	16(8)	12(9)	20(5)	19(6)	24(1)	19(6)	22(3)	21(4)	23(2)	24-12(1-9)
RR17	19(5)	15(7)	14(8)	13(9)	17(6)	24(1)	22(3)	23(2)	21(4)	24(1)	24-13(1-9)
Total (%)		6(27%)		12(55%)		12(55%)				7(32%)	

Appendix 7.12
Proficient raters' total scores and rankings in parentheses in the main study

Rater	N1(rank)	N2(rank)	N3(rank)	N4(rank)	N5(rank)	N6(rank)	N7(rank)	N8(rank)	N9(rank)	N10(rank)	Range
R2	17(4)	11(7)	12(6)	12(6)	19(3)	24(1)	21(2)	21(2)	13(5)	24(1)	24-11(1-7)
R3	15(6)	17(4)	15(6)	11(8)	14(7)	24(1)	16(5)	18(3)	15(6)	21(2)	24-11(1-8)
R6	19(5)	9(8)	18(6)	13(7)	22(3)	24(1)	20(4)	23(2)	20(4)	24(1)	24-9(1-8)
R8	16(4)	14(5)	10(8)	9(9)	13(6)	23(1)	16(4)	17(3)	11(7)	22(2)	23-9(1-9)
R14	17(3)	12(6)	11(7)	9(9)	16(4)	22(2)	15(5)	16(4)	10(8)	24(1)	24-8(1-9)
RR1	20(2)	16(6)	17(5)	16(6)	19(3)	22(1)	17(5)	18(4)	16(6)	22(1)	22-16(1-6)
RR3	14(5)	12(6)	7(9)	8(8)	17(4)	20(2)	19(3)	23(1)	10(7)	23(1)	23-7(1-9)
RR4	15(4)	8(8)	10(7)	11(6)	17(2)	18(1)	14(5)	17(2)	10(7)	16(3)	18-8(1-8)
RR8	20(4)	6(8)	13(7)	15(6)	21(3)	24(1)	18(5)	23(2)	18(5)	23(2)	24-6(1-8)
RR11	22(3)	16(5)	16(5)	14(6)	22(3)	23(2)	19(4)	24(1)	19(4)	22(3)	24-14(1-6)
RR12	17(3)	12(7)	16(4)	13(6)	16(4)	24(1)	15(5)	17(3)	15(5)	21(3)	24-12(1-7)
RR13	14(6)	6(10)	10(9)	11(8)	16(5)	21(4)	23(2)	22(3)	13(7)	24(1)	24-6(1-10)
RR15	16(5)	8(10)	12(8)	10(9)	13(7)	21(2)	19(4)	20(3)	15(6)	24(1)	24-8(1-10)
RR16	13(6)	11(7)	13(6)	8(8)	17(4)	20(3)	6(9)	23(2)	14(5)	24(1)	24-6(1-9)
RR18	18(5)	15(6)	13(7)	10(9)	20(4)	24(1)	18(5)	21(3)	12(8)	23(2)	24-10(1-9)
Total(%)		8(53%)		6(40%)		9(60%)				8(53%)	

Index

D

E

F

G

H

I

J

L

M

N

O

P

Q

R

S

T

U

V

W