Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

DOKTORI DISSZERTÁCIÓ

BENKE ESZTER

AN INVESTIGATION OF RATER AND RATING SCALE INTERACTION IN THE VALIDATION OF THE ASSESSMENT OF WRITING PERFORMANCE

Neveléstudomány Doktori Iskola
Doktori Iskola vezetője: Dr. Bábosik István egyetemi tanár
Nyelvpedagógia Program
Program vezetője: Dr. Károly Krisztina PhD.

A bizottság tagjai:
Dr. Klaudy Kinga DSc., egyetemi tanár (elnök)
Dr. Dávid Gergely PhD.
Dr. Szabó Gábor PhD.
Dr. Kontra Edit PhD.
Témavezető: Dr. habil. Kormos Judit PhD.

Budapest, 2007

Abstract

The primary purpose of this thesis is to investigate the functioning of an operational rating scale applied in the assessment of intermediate writing tasks. The validation process should ideally precede the onset of the use of such an assessment instrument, yet the need for real data defers its implementation to a later period when operational data are readily available for analysis and investigation.

The research set out to identify sources of measurement error associated with the rater-mediated subjective assessment of writing performance using two different methods of data analysis. Firstly, with the tools of modern test theory, a quantitative approach was adopted for the analysis of the assessment instrument: the six-point analytic rating scale. This investigation was further extended with the observation and exploration of rating behaviour relying on qualitative data obtained from verbal protocols and interviews that tap into the complexities of the rating process.

The validity of the rating process was established as a result of the validation of the interaction of the rating scale and the raters operating the scale. The findings seem to attest to the proper functioning of both components of the assessment procedure, the raters and the rating scale, and confirm its psychometric validity. Minor sources of malfunctioning and potential sources of non-systematic error were nevertheless detected.

The value of the research lies in the possibility of transferring the IRT method to the validation of the assessment tools used in other performance tests for which the current project might serve as a model. In addition, the practical results of the research can be incorporated into everyday testing practice with the aim of achieving the best testing practice possible under the given institutional constraints.

Acknowledgements

The following lines are meant to be more than just the compulsory tribute to people who are somehow related to the writer of the following 200 pages. Lack of space restricts me from listing all those I feel indebted to for their constant support and encouragement.

Nevertheless, there are indisputably several people who I have to express my heartfelt gratitude to, and with the help of whom I would never have been able to complete this project.

At the time of the accreditation of our examination system, my boss at the time, disregarding my obvious reluctance, involved me in the work. This is how a long-lasting devotion to testing and my career as a language tester started. Thank you, Sári.

Next, I feel deeply indebted to two experts, Barry O'Sullivan and Dávid Gergely who not only aroused my interest in IRT but also endowed me with the necessary knowledge to use for practical purposes. Special thanks to Barry for acting as a help-desk and providing on-line support during the writing of the thesis.

No research can be carried out without subjects. My special thanks to all those fifteen colleagues who provided data for my second study. I am grateful to them for acting as subjects and for doing so with such enthusiasm and cooperation.

Thanks are due to my supervisor, Kormos Judit, who, unlike me, never ever had the slightest doubt that my thesis would be completed. Her reliance urged me to try and reach her expectations. I am thankful for her help and grateful for her patience. Most importantly, I appreciate what I have learnt from her in the field of research methodology.

Erika, thank you for being the most wonderful boss in my life and for always knowing when to ask me and when to leave me alone. Thank you for allowing me to play around in the theme park of language testing, trusting me and supporting me in whatever strange ideas I came up with.

Finally, but most importantly I would like to apologize to my family for enduring the sight of my back for quite a while now, and for having to stand my waning interest in things definitely more important than my thesis. Please believe me that whatever I accomplish in my career, you will always be my greatest and most important achievement and asset.

Thank you all.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Introduction

The first chapter attempts to provide an introduction to the thesis by first briefly discussing the background to the research. Having set the aims and defined the rationale of the empirical investigation, the chapter will conclude by the presentation of the research questions that the study addresses.

1.1 Background to the research

More than a century ago, in the Journal of the Statistical Society, Edgeworth (1890) expressed his concerns about the element of chance in competitive examinations.

> Marks as measures of proficiency act like an uncorrected barometer, one in which the column corresponds only roughly, and on an average of many measurements, to the pressure of air, owing to oscillations caused by violent changes of temperature. The examinations are a very rough test of merit. (p. 460)

It is not difficult to recognize in this metaphor one of the most significant assumptions of measurement theory, namely the difference between the precision of objective measurement and the latent inaccuracy of estimation as carried out in educational testing. His metaphor of the column in the barometer translated into testing terminology is the actual observed score as captured by the measurement instrument; the pressure of the air is the target of our measurement; and the oscillations are changes in our measurement due to measurement error. Educational testing does indeed strive to

target and achieve objective measurement; nevertheless the processes involved inevitably entail error elements. Progress in educational measurement theory coupled with advances in technology have significantly contributed to the elimination of the element of chance in examinations, yet there still seem to remain factors that pose a major threat to the validity and reliability of our measurement instruments as well as measurement processes.

In language test development it is essential that each aspect of the measurement procedure should be validated. Language testers appear to attribute more attention to the validation of language tests; considerably less attention is paid to the validation of the rating process and within that the rating scale. The focus of my study is the investigation of the interaction between the rater and the rating scale, and the aim of the study is to identify elements that might contribute to the emergence of error in the measurement process. It is essentially important in high-stakes testing to identify and eliminate the factors that pose a serious threat to the validity and reliability of the measuring instrument in the evaluation process. Objectively scored tests offer less scope for biases and unorthodox rater behaviour owing to the well-defined numerical categories assigned to and applied in the evaluation of items in the test. The factors that threaten the validity and reliability of subjectively scored tests, however, are remarkably diverse and more problematic to measure. Work on subjectively scored performance tests is limited in scope and has a less extensive past than those with objective scoring. A growing interest has been apparent recently and the scope of research has been widening to include new approaches to the development and validation of subjectively scored tests.

The rationale for the research rests on the fact that the purported lack of objective scoring methods in the assessment of subjective tests increases the need for

highly valid measuring instruments and reliable measurement procedures. This is a particularly important issue in the case of large-scale high-stakes tests, most of which include subjectively scored components. The need for the ongoing validation of a testing system devolves a clear responsibility to investigate its properties, within that certain heeded aspects of the assessment procedure, to ensure highly professional decision-making based on the results of the test. Apart from test validity and reliability issues, the ethics of language testing also warrants due consideration of the approaches that improve the quality of test score interpretation and provide a more accurate picture of test-takers' ability.

1.2 The aim of the research

The direct aim of this study was to establish the validity of an operational rating scale used in the assessment of an intermediate writing task in an ESP examination. By investigating the rating process, the purpose of the research was to identify factors in the rating process that might lead to common sources of systematic and unsystematic measurement error. This was done by collating quantitative data describing the operation of the rating scale over a three-year period complemented with recently obtained qualitative data explaining rater behaviour. In diagnosing possible problems in the rater-rating scale interaction, latent sources of measurement error were expected to be disclosed. The long-term objective of the research is twofold: firstly, the results might have implications for rating scale improvement, and secondly, they might generate suggestions for rater training which, according to research findings, is frequently short-term, and only moderately effective. The results are also expected to be integrated into the continuous validation process of the testing system, and used in the

validation of the rating process of other subjectively scored components in the accredited examination system of the Budapest Business School.

1.3 Research questions

Building upon the issues discussed above, the following two generic and six specific research questions guided the investigation. The first set of questions intends to identify unusual interaction patterns in the assessment procedure, whereas the second group of questions focusing on rater behaviour aims to tap into the underlying reasons for discrepancies in the rater and rating scale interaction.

The following questions were related to the assessment instrument:

I. What kind of psychometric evidence is there for the validity of the rating scale?

1. Which assessment criteria generate bias of rater behaviour?
2. Which criteria elicit little variation in the distribution of the awarded scores?
3. To what extent is the halo effect or the cross-contamination of descriptor bands apparent in the distribution of scores?
4. Does the factor structure of the total scores confirm the appropriate functioning of the six-point analytic rating scale?

Questions related to the rater-assessment instrument interaction:

II. What are the sources of unusual rating patterns in the rating process?

1. Why do assessors exhibit different rating profiles across different domains of the rating scale?
2. What construct-irrelevant factors emerge during the application of the rating scale?

The investigation seeks to establish the validity of the rating process. In addition, the answers to these research questions provide useful information on which aspects of the rating scale amendments should be made, and suggest modifications in the assessment procedure that would contribute to a more extensively shared understanding and interpretation of the assessment criteria.

1.4 Outline of the study

Chapter 1 has introduced the broad context of the research which will be elaborated further in the subsequent chapters. The theoretical and practical aims have been outlined along with the rationale underlying the investigation. This chapter also lists the two major and the six more specific subsidiary questions that the study addresses.

Chapter 2 investigates the theoretical framework in which the study is embedded. First, some basic assumptions related to language test validation will be discussed. The second part of the chapter focuses on measurement error: from the most general theoretical approach to measurement error, the discussion will move on to its specific form prevalent in the assessment of subjectively scored tasks. This will be followed by the exploration of procedures that are commonly applied in approaches driven by modern test theory. Finally, the most influential theoretical models related to the rating process will be reviewed in order to identify elements that might emerge as sources of error.

Chapter 3 examines the empirical background of the research. The results of the most influential empirical studies investigating rater variability in the assessment of subjectively scored tasks will be reviewed. Although my focus is the rater and rating scale interaction in subjective assessment, other aspects of the rating process might also have an indirect impact on this interaction. A brief overview of validation methods will serve as an introduction to the detailed discussion of empirical research into rater and rating scale interaction in the assessment of writing performance, and a concise

summary of the application of IRT in educational science in Hungary will conclude the chapter.

Chapter 4 proposes an overview of the research methodology applied in the study. This chapter includes the description of the methods for data collection both for the quantitative and the qualitative data. Then the data analysis will be explained in more detail. As only a rather limited amount of the possible output that such analyses can yield will be used and interpreted, the description of the IRT based data analysis will also be relatively restricted owing to its rather complex nature.

Chapter 5 discusses the results of Study 1, which investigated rater effects in the evaluation of writing performance. The Many-faceted Rasch analysis casts light on how raters interpret the rating scale, and how consistent they are in the use of the six-point analytical assessment instrument. In addition to confirming rater and rating scale validity, the data are also helpful in identifying possible problems with the band thresholds as well as the four assessment criteria.

The results of the investigation of rater behaviour during the rating process are presented in Chapter 6. The quantitative results describing rater characteristics in Chapter 5 are complemented by further data on rater characteristics obtained from think aloud protocols. The role of each writing performance dimension is investigated during the rating process: the task, the performance, the candidate, the rater, the score and the rating criteria.

Further qualitative inquiries provide data for Chapter 7, in which raters' perceived behaviour is discussed based on the results of interviews. In this chapter, rater misbehaviour is explored. Participants' measured and perceived leniency and harshness are compared applying data from different sources.

Chapter 8 concludes the thesis by answering each research question. The implications of the study are discussed with a strong focus on the practical yields of the research. Additionally, the results offer ways of generalizing the findings to other types of tasks with rater-mediated assessment. The shortcomings of the research will also be highlighted besides listing further unmapped areas worth investigating in relation to the topic.

Chapter 2: Theoretical Framework

Introduction

This chapter is in four parts. First, the theoretical notion of validation in language testing will be briefly discussed which will be followed by a general overview of the most frequently applied validation methods in language testing. Next, a short description of measurement error in psychological measurement will be offered with special emphasis on educational testing. The theoretical issues will be investigated from the perspective of test validity. Besides outlining how error might contaminate the measurement process, bias, a special form of measurement error will be explored. The third part of the chapter will introduce the analytic approach, IRT that makes it possible to systematically tap into measurement error, and identify areas where the objectivity of measurement might be threatened. Finally, by linking measurement error to subjective assessment, theoretical models of rater-mediated subjective evaluation will be put forward.

2.1 Validation in language testing

Psychometricians and language testers seem to have generated an abundance of various definitions and typologies of validity. In addition, the meaning of the concept has significantly changed over time. This keen interest highlights the due importance of the issue of validity in language test development. A significant period which advanced the concept of validity was in the mid 1950s, when the American Psychological Association (1954) first offered guidelines to be followed in test development. The first version of their influential standard setting work in psychological testing, the Standards for Educational and Psychological Tests and Manuals (APA, 1966) was published more than a decade later. This was shortly followed by the influential ideas on validity put

forward by Cronbach and Meehl (1955). They made an attempt to delineate a clear path in the muddy waters that the aggregation of types of validity seems to have stirred (p. 281). They drew a distinction between four types of validity: predictive or concurrent validity, content validity and construct validity. In discussing validation procedures as early as in 1955, Cronbach and Meehl directed the attention to mathematical procedures investigating scoring that might reveal negative evidence on construct validity. They claimed that "a mathematical argument has shown that scores depend on several attributes of the judge which enter into his perception of any individual (p.289)".

The second significant period in the extension of the concept of validity is related to Messick conceptualization of validity. Since the appearance of his framework of validity (1988, 1989, 1995), a unitary concept of validity as "an integrated evaluative judgement" (Messick, 1989, p.13.) appears to be widely accepted. Three basic principles seem to underlie his framework. Firstly, special emphasis should be attributed to the interpretation of test scores with regard to its value implications and social consequences. Secondly, Messick regards validation as a continuous rather than a one-off practice. Thirdly, the accumulation of various forms of evidence should support construct validity which is regarded as an overarching concept in his framework. In line with the final claim, the present study applies the meaning of validity in a broad sense which includes aspects of validity as well as elements of reliability: the former focusing on a test measuring what it is purported to measure, the latter meaning measurement in a consistent manner. This approach seems to be justified by the fact that the interrelatedness of the two concepts is well known, "the one almost imperceptibly merges into the other" (Alderson, 1991, p.62), or as Alderson and Banerjee (2002) refer to "Messick's unitary view of validity, reliability has been merged, conceptually, into a unified view of validity" (p.102).

Whereas it is unquestionable that "validation in language assessment is ominously important, arbitrating educational and linguistic policies, institutional decisions, pedagogical practices, as well as tenets of language theory and research" (Cumming& Berwick, 1995, p.1), at the same time it is also clear that "establishing validity in language assessment is by all accounts problematic, conceptually challenging, and difficult to achieve (ibid.)". The aim of validation is "not to support an interpretation, but to find out what might be wrong with it" (Bachman, 1990, p.257). Angoff (1988) gives a list of sixteen various types of validity in his overview of the psychometric literature from 1930 onwards. Since then, however, further "validities" have emerged: Brown's (2001) decision validity and Weir's (2005) context-based validity or scoring validity are probably not the last ones to complete the list.

The analytical tools that facilitate the validation processes may be qualitative and quantitative; ideally the two types complement each other in the test development process. Various factors may inform the test developer's choice of research method, such as the type of test, sample size, or scoring methods. It is also important to consider whether the object of validation is the measurement instrument (the language test) or the assessment process (marking), or some other element of the assessment procedure (assessment scale). Besides the basic and most commonly utilized quantitative approaches of classical test theory extensively documented in the literature (Alderson, Clapham, & Wall, 1995; Bachman, 1990; Brown, 2001; Brown and Hudson, 2002; Henning, 1987; Hughes, 1989; Woods, Fletcher, & Hughes, 1986), test validation is frequently carried out with methods of modern test theory. Recent advances in quantitative test validation are especially significant in the field of performance assessment (Bachman & Eignor, 1997). Item response theory and factor analysis are the most commonly applied validation procedures in modern test theory, (Brown, 2002;

Crocker & Algina, 1986; Henning, 1987; McNamara, 1996) followed by the recently emerging method of structural equation modelling (Kunnan, 1995; Purpura, 1996). Literature discussing qualitative validation methods is less extensive probably due to its relatively novel nature, as Banerjee and Luoma (1997) argue, but verbal protocol (Cohen, 1984; Green, 1998; Milanovic & Saville, 1994; Weigle, 1994), observation (Lynch, 1996; Wall & Alderson, 1993), questionnaires and interviews (Brown, 2001; Milanovic & Saville, 1994), as well as discourse analysis (Brown & Lumley, 1997; Lazaraton, 2002) unquestionably give an invaluably revealing insight into the cognitive processes involved in language testing. This brief overview served to illustrate the most commonly applied validation methods in language testing. It seems, however, necessary to lay special emphasis on the fact that validation is not a mere investigation of a "naked instrument", as "… the instrument …is only one element in a procedure, and a validation study examines the procedures as a whole. Every aspect of the setting in which the test is given and every detail of the procedure may have an influence on the performance and hence on what is measured" (Cronbach, 1971, p. 449).

2.2 Measurement error and classical true score theory

One of the factors that has been identified as most important in tapping into potential sources of test unreliability and invalidity is error in the measurement process. The simple graphic representation in Figure 1 refers to attitude scaling. It presents the idea that measurement is contaminated with error, and the size of the error may vary.

(Oppenheim, A.N., (2004). *Questionnaire design, interviewing and attitude measurement*. p. 151.)

Figure 1 The relationship between the true measure and the error component in measurement

As Oppenheim (2004) claims, the exact nature of measurement error is unknown, what is known is that "our measure is 'impure' or 'contaminated' but not always by how much or in what way" (p. 151). It is impossible to eliminate measurement error completely, however, it is vitally important that the sources of measurement error should be explored and minimized to the greatest possible extent. Regardless of the type of measurement, each observation has an error component. The hypothetical true score reflects the real ability of the test-taker, from which any test can only take samples for the purpose of measurement. The result of the measurement of the sampled ability is the observed score which contains measurement error. The greater the measurement error, the less accurate the picture of the test-taker's ability, the less reliable the test score.

Measurement error in classical test theory is conceptualised as part of an observed scored in addition to the practically inaccessible true score.

observed score = true score + measurement error

True score theory, which postulates that each measurement has an error component was advanced in 1904 by Charles Spearman, the British psychologist. Measurement error is central to the classical true score model which, despite all its limitations, has been one of the most influential measurement models for more than a century. True score theory is also important as it is the foundation of reliability theory. Whereas true score theory postulates measurement error as one single random component, modern test theory takes a further step in the interpretation of the error element. Measurement error as posited in modern test theory can be decomposed into systematic and random error (Crocker & Algina, 1986; Nunnally & Bernstein, 1994; Viswanathan, 2005). Random error has no consistent effects; it fluctuates each time a measurement is made. Although it results in the instability of measurement, but as the total adds up to zero, it only adds variability to the data but does not change the average. Unlike random error, systematic error has a consistent effect on our measurement, either positive or negative. The observations will be stable but inaccurate, and this type of error introduces bias into our measurement. Systematic error does not affect the reliability of the test as it confounds each measurement, but it influences the validity of the measure. Non-random error as Carmines and Zeller (1979) argue, "has a systematic biasing effect … and lies at the very heart of validity" (p.14).

One special form of measurement error is bias, which is a consistent deviation in the measurement process. It is a distorting effect on the scores awarded, which is generated by the interaction between the various participants and elements of the rating process. Bias, as defined in the Standards for educational and psychological testing (APA, 1999) is "in a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct-irrelevant

components of test scores that differentially affect the performance of different groups if test takers" (p.172).

2.3 IRT in identifying measurement error

Although true score theory is the basis of most measurement applications and several attempts have been made to extend the true score model to performance rating in order to capture its subjective nature which is due to the human judgement element involved in it (Choppin, 1982; De Gruiter, 1984; Saal, Downey & Lahey, 1980), it is Many-faceted Rasch measurement (MFRM), which permits a relatively objective approach to subjective assessment. First, a short discussion of the Rasch model will be provided. This will be followed by brief, non-technical presentation of the MFRM model, which enables us to detect bias in the measurement.

Basic Rasch models are probabilistic measurement models which are primarily applied in psychological and attainment assessment and are being increasingly used in other areas, including the health profession. Item response theory (IRT) was developed to overcome difficulties associated with classical test theory. IRT does not make assumptions about sampling or normal distribution; neither does it consider measurement error to be the same for all items or persons. The basic model has a common measurement (log-linear) scale for item difficulty and person ability. The model provides estimates of each examinee's ability and each item's difficulty and locates them on a common log-linear scale. The probability of a correct response to an item is simply a function of the difference between examinee ability and item difficulty. The Rasch model, a one-parameter item response theory model, has traditionally been used for the analysis of multiple choice examinations, where the parameters involved are the difficulty of the test items and the ability of the examinees. Besides multiple

choice tasks types, other discreetly scored item-based tasks may be subject to IRT analysis. The application of Rasch analysis is inevitable in test validation, and in the standard setting procedures in defining cut-off scores. Item-banking, test equation and test calibration are also unfeasible without the use of modern test theory. There are certain factors which might distract language testers from using IRT in day-to-day language test development processes. Firstly, the relatively large sample sizes for the analyses might not be readily available in the case of tests administered to small populations. Secondly, this kind of analysis requires a slightly more sophisticated computer literacy than other procedures.

Many-faceted Rasch measurement (Linacre, 1989) is an extension of the one-parameter Rasch model which is capable of modelling facets of interest other than task difficulty and examiner ability. It is particularly useful for rater-mediated subjectively assessed performance tasks as it can identify and explore the unique features of the subjective scoring and assessment procedure. In the design of rater-mediated assessment systems, typically the following facets contribute to the rating: candidate ability, task difficulty, judge severity and the rating scale. The facets are quantified in log-odd units, or logits from the observed ratings, which are then located on a common linear scale. The psychometric validity of the assessment instrument can be investigated with the help of the analysis of the fit statistics, which will be explained and exemplified in more detail in the data analysis section. The model is capable of identifying unexpected occurrences and investigating the interaction between various facets or aspect of the rating procedure. The measurement of bias and differential item functioning is also possible with Many-faceted Rasch analysis. As with the help of MFRM, it is possible to analyse various facets of the assessment procedure, this kind of investigation is also referred to as differential facet functioning (DFF). Differential facet

functioning includes differential person functioning, differential task functioning and differential rater functioning. This method allows us to establish patterns that indicate specific types of rater errors. Although the main focus of my investigation is the interaction between the rater and the assessment criteria, all major sources of variability should be considered, as they might all have an indirect relevance to the validity of the rating scale, and might contribute to the emergence of measurement error.

2.4 Models of writing performance and sources of measurement error

Messick (1995) suggested that sources of invalidity take two forms: construct underrepresentation and construct-irrelevant variance. Construct underrepresentation indicates the extent to which a test does not include important parts of the test's proposed construct, and construct-irrelevant variance shows the extent to which a test measures something other than what it claims to measure. In the following section theoretical models of the assessment procedure will be reviewed with the intent of highlighting possible sources of measurement error. The discussion will proceed from a general model towards a framework which is more specific to writing assessment. Finally, the most common deviations that have been detected in the raters' behaviour will be enumerated.

Posing major threats to test validity, common sources of measurement error emerge from construct underspecification (Bachman, 1990, 2004), when the trait intended to be measured by the test is not adequately in the focus of the measurement as a result of under-sampling the achievement domain. A typical example is a test intending to measure one skill but in practice heavily relying on two skills instead. This happens in listening comprehension or reading comprehension tests in which a large amount of writing is involved in providing answers to the test questions. A more

striking example is when candidates are required to perform mathematical operations in a reading comprehension tests with numbers and data in the reading text. Test questions relying on background knowledge rather than source texts are also frequent sources of construct underrepresentation. A second major source of measurement error is associated with construct-irrelevant variance (Bachman, 1990, 2004; Brown & Hudson, 2002), when measurement is contaminated by factors not related to the measured language behaviour. Bachman (1990, p. 350) provides a comprehensive model to account for the sources of variance in language test scores. Personal characteristics, test method and random factors may constitute potential sources of construct-irrelevant variance. In the framework Bachman proposes, three major sources of error are identified: test method, personal attributes and random factors.

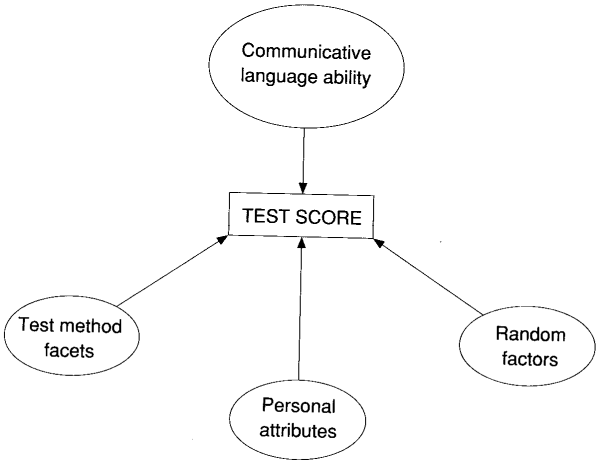


Figure 2 Factors that affect language test scores (Bachman, 1990:165)

Test score is the only element in the model which is directly accessible, and the other four elements, communicative language ability, test method facets, personal attributes

and random factors all have an impact on the observed test score. Each of these elements will have an effect of varying size on the test score and result in variation. Figure 3 gives a graphic representation of the relationship between various sources of variation in test scores.

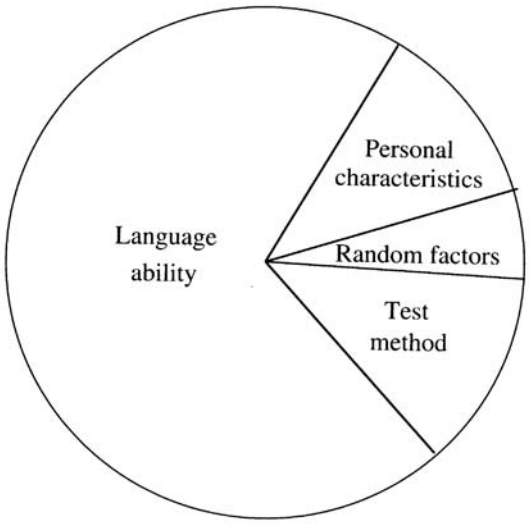


Figure 3 Sources of variation in language test scores. (Bachman, 1990, p. 350.)

The chart gives only one possible distribution of sources of variability, the proportions of these sources of variability might be different depending on a variety of factors in the measurement process. Score variance consists of true score variance and error variance, the latter stemming from uncontrollable factors in the measurement process, which threatens the reliability of the measure and therefore should be minimized. Classical true score theory has been criticized by Bachman for considering all error to be random and for failing to distinguish between random and systematic error. As Bachman

claims, systematic error might have a general effect which is one main effect constant for all observations, as well as a specific effect which is an interaction between person and facets. The systematic effect is different from random error as it introduces bias into our measures. Bachman's general model and assumptions relate to all kinds of measurement, irrelevant of the type of assessment applied.

McNamara (1996) offers a simplified conceptualization of the performance dimensions of a language test. In subjectively-scored performance testing, test method incorporates more elements that contribute to possible score variance, and within that larger error variance than in objectively scored tests. The subjective assessment of writing reflects the interaction of several elements, as shown in Figure 4: the performance that the instrument (test) elicits from the subject (candidate), the rater and the rating scale.

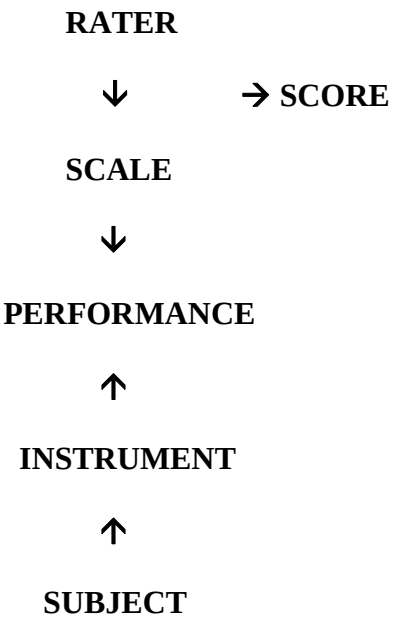


Figure 4 Characteristics of performance assessment (McNamara, 1996, p.120)

Each element of the rating process might contribute to variability and measurement error. In this model Bachman's (1990) sources of variance are presented from a different perspective, as they are more specifically tailored to subjective assessment. Although the model includes arrows indicating a possible direction of the influence, what in fact happens during the rating process is a more complex interaction (Lumley, 2005). Instead of unidirectional relationships, the process is dynamic, where the elements of the rating process enter into a complex relationship. This model is important as it highlights the element of the rating process that should be more closely scrutinized.

Whereas McNamara discusses performance assessment in general, including both speaking and writing performance, Engelhard's model (1992) is more focused on the subjective assessment of writing. In his conceptual framework, he identifies the intervening variables which, during the measurement process, provide a link between the measured proficiency and the observed rating.

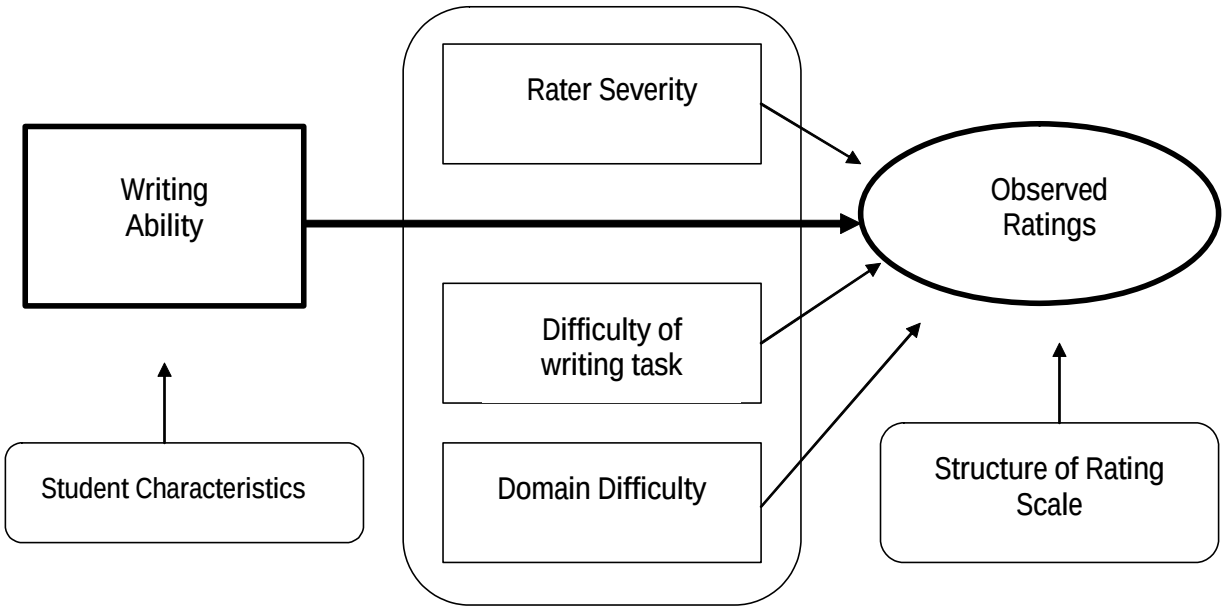
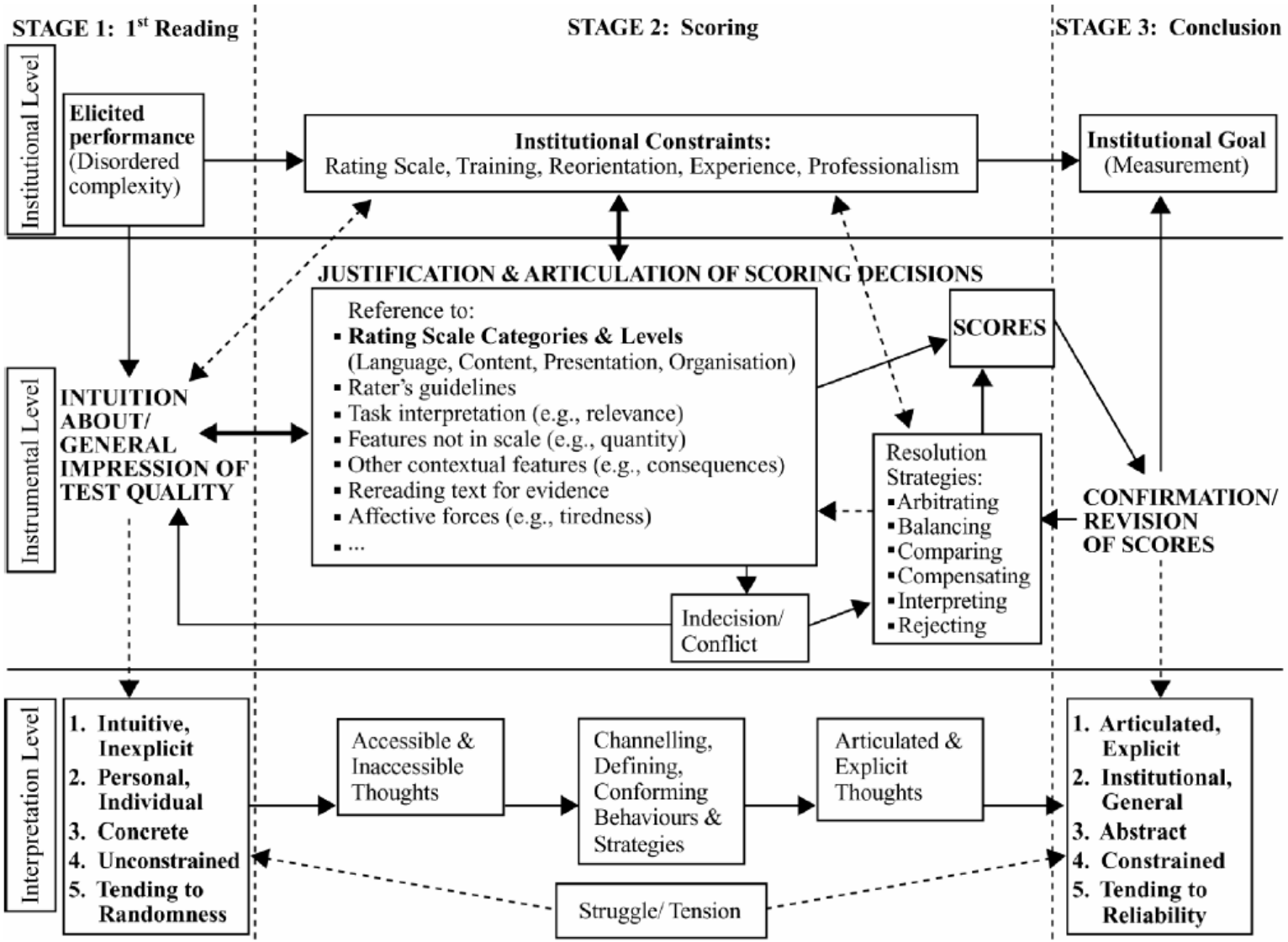


Figure 5 Engelhard's measurement model of writing assessment (Engelhard, 1992, p. 174)

The score should be a direct reflection of the ability, yet it is apparent from the model that several elements of the rating process contribute to the emergence of the final score. It is not difficult to recognize in Engelhard's model the skeleton of McNamara's framework, specially tailored to writing assessment. Student characteristics influence writing ability, but what the model fails to display is the interaction between the student and the task. In the central box Engelhard places difficulty of the writing task on one platform with domain difficulty which in fact is a feature of the rating scale. As research findings suggest in the subsequent chapter, raters are also influenced by the task, and there might also be an interaction between the rater and the rating scale. It seems that a rating process is even more interactive than what Engelhard's measurement model suggests.

A comprehensive framework for the assessment of writing performance is offered by Lumley (2005), which captures the complexities of the rating process at different levels, institutional, instrumental and interpretational level, and in three dimensions related to the stages of the rating process, first reading, scoring and conclusion drawing. His detailed model incorporates each stage of the rating process as well as various strategies raters apply in making their scoring decisions.

Figure 6 Lumley's dynamic model of the rating process (Lumley, 2005. p. 291. Reprinted with the permission of the author.)


STAGE 1: 1st Reading
STAGE 2: Scoring
STAGE 3: Conclusion
Institutional Level
Elicited performance (Disordered complexity)
Institutional Constraints: Rating Scale, Training, Reorientation, Experience, Professionalism
Institutional Goal (Measurement)
JUSTIFICATION & ARTICULATION OF SCORING DECISIONS
Instrumental Level
INTUITION ABOUT/ GENERAL IMPRESSION OF TEST QUALITY
Reference to:
▪ Rating Scale Categories & Levels (Language, Content, Presentation, Organisation)
▪ Rater's guidelines
▪ Task interpretation (e.g., relevance)
▪ Features not in scale (e.g., quantity)
▪ Other contextual features (e.g., consequences)
▪ Rereading text for evidence
▪ Affective forces (e.g., tiredness)
▪ ...
SCORES
Resolution Strategies:
▪ Arbitrating
▪ Balancing
▪ Comparing
▪ Compensating
▪ Interpreting
▪ Rejecting
CONFIRMATION/ REVISION OF SCORES
Indecision/ Conflict
Interpretation Level
1. Intuitive, Inexplicit
2. Personal, Individual
3. Concrete
4. Unconstrained
5. Tending to Randomness
Accessible & Inaccessible Thoughts
Channelling, Defining, Conforming Behaviours & Strategies
Articulated & Explicit Thoughts
1. Articulated, Explicit
2. Institutional, General
3. Abstract
4. Constrained
5. Tending to Reliability
Struggle/ Tension

Lumley uses McNamara's model as a point of departure, and expands it with his own research findings. His model is based on the results of extensive research in which he applied data obtained from both operational rating session and data generated specifically for the study. Lumley's research is immensely rich in data and aptly combines results based on qualitative and quantitative sources. Although his main focus was the rating procedure and raters' decision making processes, the analysis of the think-aloud protocols highlights significant features of the rater and rating scale interaction. Lumley concludes that although raters apply highly similar steps during the assessment of writing performances, they still have their individual interpretation of the rating criteria. Generating a score combines strict adherence to the rating scale on the part of the raters on the one hand, and an attempt to map their own intuitions on the rating scale on the other. The interaction of the elements and the processes involved in the assessment of writing tasks is clearly visualised in his pragmatic framework. The justification and the articulation of the scoring decisions in Lumley's model is informed by features constituting a direct link to the construct measured, such as the rating scale categories and levels, as well as by construct irrelevant features, including features of the performance or personal aspects of the process, for example fatigue.

From the perspective of my research the central box in the diagram merits special attention as this box subsumes aspects of the rating process which might typically result in construct-irrelevant variance. In addition, Lumley's study and the present research also share methodological similarities. Both studies apply verbal protocols to obtain in-depth data about the rating process. Lumley's study emphasizes the importance of the human aspect and involvement in the rating process and argues for the supremacy of the role of the rater in the assessment process claiming that "the validity and the reliability of the scores derived from tests depend more heavily on the

qualities on the raters (professionalism, training background, experience) than they do on those of the scale used" (p. 303).

The performance dimensions that have been proposed by the frameworks reviewed highlight sources of variability and measurement error in performance assessment which can be identified and controlled for with the application of MFR analysis.

2.5 The psychometric notion of rater misbehaviour

Finally, as the fourth component of the theoretical framework of the study, MFRM, the analytical tool for the investigation of rater characteristics and the identification of possible sources of measurement error will be discussed. The thus far identified forms of rater misbehaviour will be listed which are central in guiding the present investigation. Before discussing the most influential framework of rater misbehaviour (Linacre, 2003-2006), two other approaches will be reviewed. Saal, Downey and Lahey (1980) provide an early comprehensive summary of rating errors, in which they discuss methodological inconsistencies and the difficulties that the lack of a commonly accepted definition causes in the empirical investigations exploring the psychometric qualities of rating data. In a meta-analysis of 20 articles, they concentrate on three major forms of rater behaviour: halo, leniency or severity and central tendency or restriction of range. They also include interrater reliability or agreement in the discussion of the quality of the rating, but at the same time they seem to accept the view that this aspect of the rating is more concerned with rater reliability than with the validity of the rating. A high interrater reliability, as they claim, is not necessarily a sign of valid and accurate rating. A score might be highly reliable but perfectly invalid at the same time. Their conclusion is that clear and commonly accepted definitions are needed

for rating characteristics and consistency in the application of statistical methods to quantify them is important. “Failure to pursue alternative approaches … will surely sentence rating research to many more years of inconsistency and confusion” (p. 426).

An alternative approach to the investigation of rating behaviour which Saal, Downey and Laney called for in the early eighties emerged some time later with the appearance of Many-faceted Rasch analysis. Engelhard (1994) promotes the measurement model developed by Linacre (1989) to investigate performance assessment and provides an empirical example to illustrate the rater errors Saal, Downey and Lahey (1980) identified. He set up four error categories in handling central tendency and the restriction of the range of the scores separately. The analysis he conducted included 264 compositions and provided examples for each type of rater error. Engelhard’s conclusion is an acceptable stance in adopting a realistic approach to rater misbehaviour. He quotes Guilford (1936) when he claims that “raters are human and they are therefore subject to all the errors to which humankind must plead guilty” (p. 272), and adds that the ongoing quality control procedures significantly enhance the possibilities of rater effects to be minimized.

Many-faceted Rasch measurement (Linacre, 1989) is capable of detecting rater effects. Six forms of “known rater misbehaviour” have been identified in Linacre’s (2003-2006) MFRM investigations as exerting a distorting effect on rating. It is important to note here that Wolfe, Moulder, Bradley and Myford (2001) go even further in labelling rater discrepancies: while discussing rater effects over time, they identify similar categories to those of Linacre’s, and name them as rater aberration. The leniency and severity aspect concerns the tendency to over or underrate performance as a result of the unique personal characteristic of the rater. This form of rater misbehaviour is one of the most commonly researched areas in language testing.

Awarding scores either in the middle region of the rating scale, or, alternatively, showing a preference towards extremes is labelled by Linacre as central tendency and extremism. The third misbehaviour relates to the effect one dominant criterion exerts on the other criteria in the analytic rating scale. The halo effect is not unique to the rating process in language testing and has its origins in psychological research early in the last century (Thorndike, 1920). Two manifestations of rater misbehaviour, response sets and playing it safe, refer to the assessors' detachment from the rating scale by either using regular score patterns irrelevant of the abilities assessed or relying on other raters' evaluations, assigning ratings that agree with other raters' judgement. Inconsistency is a negative feature of the marker which is probably the most detrimental, and the most difficult to control for. Whereas the former types of rating discrepancies can be relatively well delineated and explained, inconsistencies incorporate innumerable and indefinable sources of error. These six misbehaviours, leniency and severity, extremism and central tendency, the halo effect, response sets, playing it safe and inconsistencies will be addressed in the study both by qualitative and quantitative means.

2.6 Conclusion

Rater variability and diverging rater characteristics raise concerns regarding the validity and the reliability of the measurement procedure in rater-mediated assessment. In order to identify factors that might distort the rating process, it is necessary to carry out validation studies to ensure that the effects of construct-irrelevant features are minimized. Whereas classical test theory treats measurement error as one entity, modern test theory is capable of detecting error components related to each constituent of the assessment procedure. Many-faceted Rasch measurement with the help of the

FACETS program provides a comprehensive probabilistic framework and an analytical tool to detect rater error in performance assessment.

The validation of the subjectively scored writing performances should entail the validation of each component of the measurement process. Various theoretical models exist which demonstrate the complex relationship between the candidate, the rater and the performance. It appears from the models that the assessment is a highly complex procedure in which the final score is the result of the interaction of numerous factors. The main focus of my validation study is the rater and rating scale interaction. With the application of Many-faceted Rasch measurement, the psychometric qualities of the rating scale will be investigated which cannot be viewed in isolation from the other component of the rating process.

Chapter 3: Empirical Background

Introduction

Having reviewed the theoretical underpinnings of the current study, in this chapter I will consider empirical research carried out in the area of rater-mediated subjective assessment of writing. First, as the study has a strong psychometric focus and makes extensive use of Many-faceted Rasch measurement, the chapter will start with a brief account of studies discussing how the method can be used in language test validation and rating scale construction. In the second part I will consider Rasch-based research more directly linked to the present investigation; studies will be reviewed in which Many-faceted Rasch measurement was applied to examine rater effect in the rating process. Research investigating rater characteristics, the rating process, rater biases and the effect of rater training in removing rater error will be surveyed. Finally, with a special focus on Hungary, theoretical discussions of item response theory and empirical studies applying the method as an analytical tool in educational research will be briefly reviewed.

3.1 Many-faceted Rasch measurement in language testing: rating scale construction and validation

One of the most intriguing areas to investigate in writing assessment is the rater-scoring procedure interaction. O'Sullivan and Rignall (2001) assert that "rater variation is potentially the Achilles heel of performance testing, as it represents a significant source of construct-irrelevant variance" (¶. 2). It seems a plausible argument that "the reliability of the rating scale is critically dependent on the raters who operate it" (Congdon & McQueen, 2000, p. 163). Connor-Linton (1995) rightly claims that

"research on the rating process can address many aspects of the overarching question of rating scale validity" (p. 764). Many issues related to subjective assessment require going beyond the techniques of traditional data analysis. Although the application of item response theory in psychology has a tradition of more than a hundred years (Hambleton & Swaminathan, 1985), the type of IRT model that is gaining general acceptance and is becoming more widespread in language test development can be traced back to the 1960s. Many-faceted Rasch measurement is a fairly recent addition to the existing IRT techniques. MRFM can deal with complex rating systems and has the capability of quantifying deviations from an expected model. With the help of MFRM, also called the FACETS model, it is possible to put various aspects, "facets" of the measurement process into one single framework of reference and investigate the delicate interaction between these different facets. Advances in computer technology in recent years have eliminated the need to deal with the mathematical difficulties implied by the use of the approach and made the method accessible to a wider research community. Despite the major drawback to the use of IRT, namely the fairly large sample size required for the analyses, it has proved to be an invaluable technique in numerous aspects of test validation and use (Bachman & Eignor, 1997; McNamara, 1997; Pollitt, 1997).

IRT is a commonly utilized tool to develop and validate rating scales of performance tests (Barker & Hawkey, 2004; McNamara 1996; Milanovic, Saville, Pollitt, & Cook, 1996; Shaw, 2004a, 2004b), and has received special attention recently in the ongoing development of the Common European Framework of Reference for Languages (North, 2000; North & Schneider, 1998).

Among the numerous methods the Common European Framework of Reference (CEF) lists as possible tools for scale development, item response theory is the twelfth

in line. Although the description of how the Rasch model can be applied does not provide easily comprehensible guidelines to rating scale construction, the concluding remarks in the section on IRT deserve attention. The authors stress that "Rasch can be also used to analyse the way in which the bands on an assessment scale are actually used. This may help to highlight loose wording, underuse of a band, or overuse of a band, and inform revision" (Council of Europe, p. 211). Given the apparent popularity and increasing influence of the CEF, it can be assumed that the warning that rating scales need revision reaches representatives of the language testing community.

Unlike the CEF itself, one of its authors (North, 2000) gives an extensive demonstration of how FACETS can be used for rating scale construction. Besides giving a detailed theoretical background to the scale construction project, North (2000) meticulously describes how the descriptors for the common framework scale of language proficiency were calibrated. The combination of methods of item-banking with judgemental data yielded an illustrative descriptor bank, which, as the author points out, should be submitted to further modifications. Whereas the work is impressive in its size and transparency, North (2000) does not fail to mention some problems associated with his project. Some of these shortcomings are related to the rating scale model that FACETS is based on. Nevertheless, he also adds that the ramification of these problems are not necessary the tasks of applied linguists.

Tyndall and Kenyon (1996) describe how a newly developed holistic rating scale was validated with the help of FACETS. In the first phase of the project, the scale development was mainly qualitative, intuitive and judgemental. Selected experts were asked to validate the draft scale against course expectations as stated in the faculty syllabus. The second study describes how the data obtained from the first operational use of the scales were analysed with Many-faceted Rasch measurement. Raters, essays

and the rating scale were all submitted to FACETS analysis. The findings confirmed that the different levels of the scale were clearly separated. The results also attested to the validity of the new seven-point rating scale, and provided further data on rater misfit. The authors attributed this ideal operation of the scale to two factors: firstly, raters took part in the construction of the scales, and secondly, the scales were partly constructed with the raters' internalized rating experience.

The Cambridge ESOL ongoing research agenda for the validation of second language writing assessment (Shaw 2002, 2004a, 2004b) investigated questions similar to the ones explored by the present study on a much larger scale. The ESOL research monitored the results of the changes implemented during the revision of the assessment scale, and the ongoing nature of validation was emphasized during the course of the research. This comprehensive study describes how the new assessment criteria and band scale descriptors were developed for the revised IELTS writing task. A wide variety of data collection methods, including interviews with stakeholders and questionnaires administered to assessors worldwide, were used to inform the initial stage of the development of the assessment scale. The second phase entailed the trialling of the assessment instrument with expert raters. MFRM as well as G-theory was used for data analysis in the validation process to investigate both the properties of the newly-devised assessment instrument and rater characteristics in operating the rating scale. Not only the validity of the new assessment scales was confirmed, but the ample data obtained during the validation procedure is also made use of in standardization, rater training and test development processes.

Item response theory is also referred to as a useful quantitative method in developing rating scales for speaking (Fulcher, 2003; Luoma, 2004) with the fair reservation that the large sample sized needed is not readily available in every language

testing context. Unlike the researchers referred to earlier, Milanovic, Saville, Pollitt and Cook (1996) used another IRT program, BIGSTEPS to validate a new rating scale for speaking. The probability curves for the rating categories demonstrated that raters were able to differentiate between different levels of the scale fairly well. Adding further support to North's (2000) finding, the authors in this study also identified problems with the end points of the scales, which is an issue that should be addressed as the validation of the scales continues.

All the studies surveyed in this section emphasize the need for a quantitative approach to rating scale construction. Additionally, they all underline the importance of continuous validation. It is not only the rating scale that is validated with the help of the analyses described in the studies, but the other elements of the rating process, the rater and the task are also investigated in terms of their proper functioning. In my research, Study 1 will examine the functioning of the rating scale applying methods similar to the ones used in the reviewed studies.

3.2 A many-faceted approach to the investigation of rater behaviour: rater characteristics

Apart from rating scale analysis, the FACETS model is capable of providing data about raters. Numerous rater characteristics can be disclosed based on the results of the analysis, some of which might contribute to construct-irrelevant variance and measurement error. In one of the first MFRM applications to investigate rater characteristics, Saal, Downey, and Lahey (1980) identified the following five common rater errors: severity or leniency, the halo effect, the central tendency effect, the restriction of range effect and inter-rater reliability or agreement. Rater error is also known as "aberrant rating" (Wolfe, Moulder, Bradley & Myford, 2001), especially

when investigating rater effects in differential rater functioning over time (DRIFT). Linacre (2003-2006) lists six critical rater-related issues resulting in error variance in subjective assessment. The "known misbehaviours" include rater strictness and leniency, extremism and central tendency, halo/carryover effects, response sets, playing it safe, and instability.

3.2.1 Rater leniency and strictness

Rater leniency and harshness are the most frequently investigated rater characteristics. In discussing Many-faceted Rasch measurement applications in writing assessment, and in terms of the focus of my project, three seminal studies need special attention. Engelhard (1992) provides a detailed, comprehensible justification of why the Rasch-based FACETS approach has an unchallenged supremacy over other methods in the investigation and the validation of the measurement of writing ability. The measurement model that he proposes for writing and on which his empirical investigation is based has already been referred to in the discussion of the theoretical framework of my study. To illustrate his initial theoretical reasoning concerning the FACETS measurement model, he analyzed the analytic scores of one thousand writing performances on four dimensions: writing ability, rater severity, writing-task difficulty and domain difficulty. The rating scale use and the appropriate functioning of the rating criteria were also the subject of the investigation. Besides identifying patterns in rater behaviour and quantifying the level of difficulty of the tasks used, the researcher also investigated which rating categories were harshly and leniently scored. With such a detailed analysis of the various facets of the assessment procedure, Engelhard's aim was clearly to demonstrate the advantages of the FACETS approach over other measurement models. Using FACETS to generate rater profiles in order to create rater

banks in the manner item banks are made is an idea earlier not promoted by other researchers.

In a later study Engelhard (1994) investigated four rater errors, severity, halo effect, central tendency and the restriction of the range of the awarded scores on 264 compositions, marked by 15 readers. The results of the FACETS analysis indicated significant differences in rater severity. Halo effect, that is, one rating category affecting the others, was also detected, as well as a drift to central tendency. Less apparent, but also an observable feature was the restriction of the range of the scores. In this fairly early study applying Many-faceted Rasch analysis, the author clearly showed how detailed information could be obtained about the quality of rating with this measurement model. He underlines the importance of gathering theory-driven data about rating performance for which Multi-faceted Rash measurement appears an indispensable tool.

More recently Eckes (2005) used FACETS analysis to investigate rater effect in the TestDaF speaking and writing tests. Scores obtained from the assessment of 1359 writing performances and 1348 oral proficiency interviews were analysed, with 29 and 31 raters rating the performances subsequently. The findings seem to confirm results of former studies (e.g. Engelhard, 1992, 1994; Kondo-Brown, 2002; Lee & Kantor, 2003; Lunz, Wright, & Linacre, 1990; Myford, Marr, & Linacre, 1996; Weigle, 1994, 1998) which indicate differences in rater severity. In spite of the noticeable differences in severity, however, raters displayed a relative consistency in their overall ratings. A lower level of consistency was detected though in raters' attitude to the rating scale. The global impression criterion was the one which raters treated with unusual generosity, and it was the linguistic realization criterion on which raters were the strictest.

Wolfe (2004) examined rater accuracy, leniency and centrality in the assessment of writing tasks. In the research design where 101 readers marked 28 writing tests, substantial differences were found in the level of readers' severity. Hypothesizing a theoretical cut-off point in his data, Wolfe shows that with such a criterion-level the most lenient rater would pass 75% of the candidates, whereas the most severe rater would pass only 15% of the examinees.

In a pilot study, I applied MFRM to investigate rater characteristics (Benke, 2004). The aim of this small scale project was to tap into some of the major differences between classical and modern approaches investigating rater reliability in the analysis of subjectively scored test performance. The focus of investigation was the rater-score interaction. Firstly, the double scoring of a writing assignment was analysed with the help of the statistical procedure Spearman rank order correlation. Then an IRT-based approach was applied on the same dataset. The results suggested that an MFRM analysis was far more informative about the rating process both from the perspective of test-takers and that of raters. Furthermore, it provided data about rater leniency and harshness, marker consistency and inconsistency. The results obtained with the help of the latter analysis can inform test development, and may have implications for rater training besides prompting procedures to increase the validity of the test and the reliability of the measurement.

Research on rater leniency and strictness is also concerned with the stability of rater characteristics over time. Myford, Marr and Linacre (1996) used FACETS to investigate rater characteristics in the writing assessment of TESOL tests and explored how differences in rating can be eliminated. In their research they found that although raters differed in their degree of severity, the effects were minor, only half-point adjustments seemed to be justified on the raw scores to compensate for the differences

in severity. They argued that although those half points might not be relevant for the majority of examinees but absolutely crucial for those close to the critical cut-off scores. In comparing the consistency of rater severity over time they found only modest correlation (r= 0.3) across marking periods. This result serves as a warning for test developers to fine-tune their rating designs.

Confirming Myford, Marr, and Linacre's (1996) finding, according to which rater leniency and severity are only slightly consistent over and across rating periods, Lee and Kantor (2003) conclude in their study investigating 970 writing TOEFL samples that raters displayed the highest degree of consistency in terms of severity and leniency with the easiest task and their level of severity was rather varied with the most difficult task.

In Congdon and McQueen's (2000) study, the stability of rater severity over an extended rating period was investigated. 8285 writing performances were analysed with Many-facet Rasch analysis. Significant differences between raters were found both in terms of marks awarded on one day and marks given during the whole period. These findings also confirm earlier findings according to which monitoring rater consistency is of utmost importance in performance testing.

In investigating rater severity, O'Neill and Lunz (2000) found that raters' history is a good predictor of their future performance. Furthermore, the consistency of a rater's past performance may serve as a good basis for selecting him or her as a main or chief rater. Their results are in line with the majority of the findings related to rater severity and leniency, namely that even the most consistent markers may vary in their characteristics at times.

Research carried out in the field of rater severity and leniency attest to the existence of noticeable differences between raters. The majority of the findings seem to

suggest that severity and leniency are relative characteristics. They might be highly stable features, little susceptible to considerable change within one rating period and with regard to one task, yet they are still dependent on the task, and raters might show different profiles in their strictness across rating periods.

3.2.2 The rating process

In most studies the questions that address the sources of rater variability are presented in a unidimensional view, either from a qualitative or a quantitative perspective. One exception is the body of research focusing on the rating process. In order to tap into the cognitive processes that take place during the rating procedure, investigations are mostly carried out with qualitative methods. This approach facilitates a better understanding of the rating process to address the problem which Connor-Linton (1995) presents, "if we do not know what raters are doing (and why they are doing it), then we do not know what their ratings mean" (p.763).

The ways in which markers arrive at the final score is also widely discussed in the literature. The more complex the rating scale, the more varied its interpretation and the more possible that construct-irrelevant features creep into the rating process, as Orr (2002) found in his study of FCE speaking tests. Even if examiners awarded the same scores, in their assessment procedures they put emphasis on different aspect of the performance according to the results obtained from the verbal protocols. Orr identified 10 categories, "non-criterion information" in the assessment procedure. Although he investigated the rating of spoken performance, two of his categories deserve special attention as these might also be typical in assessing written production: references to global impressions and comparing the candidate with another.

Cumming, Kantor, and Powers (2001) specified three purposes which had motivated their research into the rating process. Firstly, a thorough understanding of the rating behaviour may inform test development, validation and the development of scoring schemes. Secondly, it has fundamental implications for rater training, and thirdly it helps provide valid results to stake-holders. They sought to find an answer as to what types of decisions are made by raters and also in what sequence the decision making occurs. Three findings are important from the perspective of my study. First, the think-aloud results seem to suggest that raters treat different criteria differently: whereas native-speaker raters paid even attention to all rating criteria, ESL/EFL trained raters tended to focus on the grammar criterion. Second, it appeared that rating-irrelevant comments included raters' evaluative judgements of the task. While marking the scripts, raters tended to make comments related to the task and attributed shortcomings of the scripts to the task rather than to the writer. Third, in some cases raters tended to make comparisons between candidates and script, although the test was expected to be marked against assessment criteria. The study indicates how data collected with concurrent verbal protocol during the assessment process can inform test construction and more importantly rating scale development projects.

Vaughan (1991) also adopted a qualitative perspective in her investigation into raters' thought processes during the marking procedure. Her focus was on holistic rating, and she investigated salient features of the text and significant rater characteristics which influence the scoring. The data she obtained from verbal protocols confirm her hypothesis according to which individual approaches dominate the assessment procedure. She identified rater strategies in which one characteristic feature guides the rating. "The first impression dominates", "the laughing rater" or "the

grammar-oriented rater" all highlight aspects of rater behaviour which might divert the assessor from the rating scale.

The results of the studies investigating the scoring procedure offer insight into raters' decision-making processes. The data, however, should be treated with caution: in the interpretation of the results it should be borne in mind that these data can be obtained in experimental circumstances, mainly with the help of verbal reports, and thus may be different from what is going on in an operational rating session. For all its limitations though, verbal reports, as Lumley (2005) also suggests, is a methodology capable of producing data no other method would be capable of. The qualitative inquiry of Study 2 in my research also makes use of this data collection method.

3.2.3 Bias analysis

The studies surveyed so far confirm that rater-mediated assessment is a highly complex process in which raters all have their unique rating patterns. One special form of measurement error is bias, a systematic deviation in the measurement process. It can be conceptualised as any construct-irrelevant source of variance that results in systematically higher or lower scores. Bias research has not developed parallel with the appearance of IRT methods: interest in the notion of bias originates with the appearance of intelligence testing. The definition of bias has changed over time: whereas the earliest body of research concentrated on special ethnic groups disadvantaged by intelligence testing, by now the notion of bias has broadened.

> Measurement bias is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable groups. When evidence of such deficiencies is found at the

level of item response patterns for members of different groups, the terms item bias or differential item functioning are often used (APA, p. 74).

The aim of bias research is to facilitate fair testing practices and to ensure that no groups of gender, native-language, ethnic or socio-economic status are advantaged or disadvantaged. This principle should underlie all testing practices in order not to fall into the trap Darwin (1845) prophesized some two centuries ago: "If the misery of our poor be caused not by the laws of nature, but by our institutions, great is our sin".

The most frequently researched areas are gender-based differential item functioning (Buck, Kostin, & Morgan, 2002; Stoneberg, 2004; Takala & Kaftandjieva, 2000), ethnic-based differential item functioning (Chen & Henning, 1985; Elder, 1996; Kim, 2001; Kunnan, 1990; Sasaki, 1991; Ryan & Bachman, 1992) and culture-based differential item functioning. The majority of recent bias analytic studies apply IRT methods; some use the Mantel-Haenszel D-DIF statistic to identify potential sources of bias.

The writer facet, as a potential source of bias in writing assessment, is identified as an issue of concern in large scale high-stakes testing (Cumming, 2002). As ethical considerations bear special relevance in ensuring fairness and consistency to a wide population of candidates, potential sources of ethnic or culture-based bias are increasingly monitored by test providers. Lee, Breland, and Muraki (2005) report on their findings concerning the writing tasks of the computer-delivered TOEFL CBT examination. Although they found significant group effects in one third of the total 81 prompts examined, the effect size was too small to consider it particularly important and concluded that writing prompts are comparable for examinees of different native language groups.

Besides studies investigating how special groups might be disadvantaged by the measurement instrument, it is also interesting to explore what forms of biases raters display in the subjective assessment procedure. O'Sullivan and Rignall (2001, 2002) examined rating characteristics in a longitudinal study applying data from bias analysis. Sixteen raters took part in the project, and rated sixty scripts of the IELTS General Training Writing Module in a linked rating design. The rating was organized in four rounds. The bias analysis applied in the study modelled unusual rating patters, identified differential rating criterion functioning and provided data on inter-rater and intra-rater reliability. The MFR analysis of the rating criteria investigated over a period of four rating sessions revealed that out of the five assessment criteria only one seemed to be applied consistently, this was the harshest criterion, that is, the one which raters scored strictest on. There was some variability in the use of the other criteria but within a much smaller range of logit values. Although the bias analysis identified unusual rating patterns but there "did not seem to be a significant trend in the data" (p.17).

In a seminal paper, Kondo-Brown (2002) explored rater bias in the assessment of norm-referenced Japanese L2 writing. The aim of the investigation was twofold: firstly, it analysed rater bias towards certain assessment criteria in evaluating second language writing performance, secondly, it attempted to identify special groups of candidates towards which bias could be detected in an in-depth analysis of the candidate-rater interaction. The FACETS analysis of the data showed that there were differences in terms of severity among the raters as well as in their treatment of the rating criteria. It should be noted, however, that although these differences were minor in terms of logit values, the differences were still significant. The reliability of the separation index also confirmed that the raters were not equal in terms of their harshness. The bias analysis also revealed slight biases, especially towards one of the

prompts. This was the task given to the candidates on the first day, thus it might be hypothesized that the additional rater training before marking further writing tests on the second and third day might have removed the sources of biases. The author stressed the need for further analysis to explore whether it was training that removed the bias associated with the prompt which was marked in the first rating session or whether the bias should be attributed to some different source related to the prompt itself. What is especially interesting in the findings is that each rater showed a unique rater-candidate bias pattern. A similar conclusion could be drawn from the rater-criteria interaction: although biases were detected, there was no general pattern of rater bias towards any of the rating criteria. In line with other findings concerning rater bias, Kondo-Brown (2002) confirmed the highly idiosyncratic nature of assessors' rating behaviour and proposed further qualitative investigations.

The results of the studies investigating differential criterion functioning indicate that although bias terms can be identified with the help of MFRM, there does not seem to be in a trend in raters' awarding systematically higher or lower scores to candidates on any of the assessment criteria.

3.2.4 Rater training

So far rater characteristics have been described with the help of studies carried out to investigate rater idiosyncrasies. In order to create a more extensively shared understanding of the evaluation system of the examination as well as to eliminate inconsistencies, rater training appears to be an essential tool. Numerous studies have been conducted to investigate the effect of rater training. Research findings related to its practical use, however, are slightly controversial. Besides the logically anticipated findings, namely the usefulness and effectiveness of rater training in enhancing rater

performance and consistency (Weigle, 1994, 1998, 2002), there also seems to be empirical evidence to suggest that the effects of rater training are short-term, and certain differences between raters cannot removed by the training (Lumley & McNamara, 1995; Wigglesworth, 1993).

Lumley and McNamara (1995) investigated the stability of rater characteristics. As regards the efficiency of rater training, they concluded that there "is a substantial variation in rater harshness, which training has by no means eliminated, nor even reduced to a level which should permit reporting of raw scores for candidate performance" (p. 69). They also carried out bias analysis to find out how rater harshness changed over time. Like other researchers, they also stress the importance of intra-rater stability and see the purpose of rater training in enhancing self-consistency.

Differences between experienced and novice raters were examined by Weigle (1998) in a study which compared pre-training and post-training rater reliabilities. The Many-faceted Rasch analysis indicated that novice raters were stricter and less consistent than their more experienced colleagues. Both groups received training after which the differences between the two cohorts were less marked than prior to the training. Novice raters' consistency showed considerable improvement, while differences in strictness still remained. The FACETS analysis suggests that rater training is more effective in terms of increasing intra-rater reliability, that is, rater consistency, rather than in changing strictness or leniency.

The effect of feedback to raters on their rating practice is also an issue of concern. In investigating the assessment of oral interaction tests, Wigglesworth (1993) looked into the effect of providing feedback in the form of an "assessment map" to the raters on the consistency and severity of their rating. The bias analytic approach allowed identifying idiosyncrasies of the rating patterns. Presenting the results of the

FACETS analysis to the raters resulted only in a slight improvement of the rating performance.

Immediate feedback on rating characteristics might be a valuable tool in eliminating rater effect, as Wilson and Case (2000) suggest. Even training may fail to eliminate major differences in harshness, which, according to their results may display substantial differences across rating periods. Although information from psychometric analysis does improve the quality of single rating, double or triple rating substantially contributes to the reliability of the awarded grades, as the authors suggest.

The effect of feedback based on bias analysis was the focus of the research carried out by Elder, Knoch, Barkhuizen and von Randow (2005). They were cautious in attributing the slight improvement of the ratings to the feedback only, but nevertheless noted raters' positive attitude to receiving feedback on their work. Elder et al. came to the conclusion that "feedback appears to have come at a price" (p.190), and as a result of more homogeneity in the rating the discriminatory power of the tests apparently decreased. They were more tentative than O'Sullivan (2006), who "found unpredicted consequences to feedback, and several of the participating raters swung wildly from one extreme to another as they tried to include … feedback in their rating" (p.89). On the whole O'Sullivan (2006) seems to share the opinion of those who are doubtful about the unquestionable effectiveness of sharing rating characteristics with the markers.

It seems from the studies that opinions differ as to the efficiency of rater training and feedback given to raters on their rating performance. There is an agreement between researchers, however, that by reducing random error, training improves raters' self-consistency and eliminates extreme differences between raters. Thus reducing variability in rater behaviour might result in more valid and reliable scores.

3.3. IRT in educational measurement in Hungary

IRT applications have been used increasingly in educational measurement in Hungary in recent years. In general educational science and psychology, both the theory itself and its field of possible application are well documented both in Hungarian and in English. In his earliest work on measurement theory, Horváth (1985; 1991) gives a clear introduction to the Rasch model in a comprehensive account of the development of psychometric theory. A detailed discussion and comparison of classical test theory and modern test theory constitutes the focus of Horváth's work (1993), which caters for the needs of those who would like to grasp the complex mathematical explanations on which the deterministic and the probabilistic theory is built. Although the author explicitly recommends this book for readers with a strong background in science, language testers with a less profound understanding of the mathematical complexities might find his work highly instructive. A highly technical description of modern test theory is presented in Horváth's more recent book (1997), in which he illustrates the theoretical claims with practical examples to render the operation of this sophisticated model more comprehensible. He also extends the possible use of the method to questionnaire data analysis (Horváth, 2004) and briefly mentions how the IRT based rating scale model can be applied in questionnaire data analysis.

Csapó (2004) discusses the basic concepts of this "new generation of modern, probabilistic test theories (p.282)" in his succinct but highly informative introduction to proficiency testing. In exemplifying and interpreting item characteristic curves, he points out the difference between the deterministic nature of classical test theory as opposed to the probabilistic nature of measurement based on modern test theory. He maintains that in educational testing one of the most useful of the family of probabilistic models is the two-parameter Rasch model.

Empirical research using IRT in Hungary is also notable in the field of education. The differences between two IRT-based computer programs, OPLM (Verhelst, Glas & Verstralen, 1995) and ConQuest (Wu, Adams & Wilson, 1998) in the assessment of teenage schoolchildren's reading, mathematics and science skill are discussed by Molnár (2003). The main aim of the study was to promote the use of IRT applications in educational measurement and popularize the terminology associated with this fairly uncommon, although not altogether new field of study. The author gives additional evidence in further studies how the use of the Rasch model can assist educational research (Molnár, 2004; 2005; 2006).

In the Hungarian language testing literature, Bárdos (2002) compares classical test theory to modern test theory, and discusses how language testing can benefit from the application of probabilistic methods in the assessment of language proficiency. In a highly comprehensive discussion of the relationship between person ability and item difficulty on which the theory is premised, he illustrates how the chance or the probability of the candidate answering an item correctly is a function of the person's ability. The author also discusses the differences between the one-, two and three parameter models, and completes the explanation with warnings concerning the required sample sizes for the analysis to obtain stable results. The current study attempts to challenge his slightly pessimistic conclusion, which, although acknowledges that psychology has introduced the methods applied in modern test theory into language testing, but at the same time maintains that " in everyday practice mathematical inventions and innovations … remain unnoticed" (p.50).

Besides theoretical work dealing with IRT in the field of foreign language testing in Hungary, two major empirical studies merit attention, which are prime responses to the challenges modern test theory offers. Dávid (2000), in his unpublished

PhD thesis focuses on the validation of two test methods. In both studies he investigated task types which are relatively infrequently applied in language testing. Whereas in Study 1 he investigated the multitrak item types of the former entrance examinations, in Study 2 he concentrated on a fairly special oral examination format: the small group orals. In search of systematic error associated with the task formats in issue, the author provides a detailed description of item response theory including the basic as well as the extended model. The study provides evidence on how item response theory together with other, both qualitative and quantitative validation methods can inform the test development process. It is important to bear in mind that "even in objectively scored test components, method is far from neutral unless its effects are compensated for" (Dávid, 2007, p. 94). Study 2 includes references to the development of the arsenal of analytic tools used by the author's institution and what further paths to test development were opened up with the help these tools. In the context of discussing orals, Dávid gives examples how FACETS, the computer program used for Many-faceted Rasch measurement can be applied in the field of subjectively scored tests, and exemplifies further aspects of test performance that can be investigated with its help. References are also made to the potentials of modern test theory for the creation of item banks.

How an item bank can be constructed with the help of IRT is clearly illustrated by Szabó (2000) who presents the example of a six-year project in his PhD thesis. With a detailed account of fundamental psychometric concepts which informed the test development process, the author gives an insight into the laborious work required for the creation and the operation of an item bank. In providing the theoretical underpinnings of the practical work carried out, the author takes an unbiased approach: besides presenting the numerous advantages modern test theory can offer over classical test theory, he does not fail to mention the often voiced criticism against IRT. He

rightly concludes that IRT is not an omnipotent tool to overcome all problems classical test theory cannot cope with; it is rather a novel approach to complement and to fine-tune existing psychometric knowledge. He also adds that there are areas which can definitely benefit more from IRT than others; and item banking is one these fields. The success of the project described and the account of how the operation of the item bank ought to be monitored (Szabó, 2006) encourage other test providers to follow their example even if the expertise required and the resources available to most of them are far from what would be needed.

3.4 Conclusion

This chapter sought to review the most important empirical studies which investigated sources of variability as potential sources of measurement error in the assessment of writing performance. From the perspective of my research an important aspect of investigating the validity of the rating scale is concerned with the rater and rating scale interaction. The studies reviewed in the earlier part of the chapter highlighted how IRT applications can be used in language testing, especially in rating scale construction and validation. The latter part of the chapter emphasized the valuable difference the use of IRT can bring to the investigation of rater variability and the validity of the rating. Many-faceted Rasch measurement makes it possible to analyse the psychometric qualities of the rating process: those of both the rating scale and the raters. Studies applying bias analysis, a special method to identify unmodelled rater variation were also discussed.

Despite the breadth of existing research on the assessment of writing, it appears that there is still scope for more research that is based on real data collected from several operational rating sessions, and which looks specifically into sources of rater

misbehaviour. Previous research findings have informed my research design, and will facilitate the interpretation of the results. To establish the validity of the rating process, this study will also investigate the properties of the rating scale and the behaviour of raters using MFRM. Unlike earlier research, this validation process was carried out on a rating scale used across different languages. In addition, the present study took a stronger qualitative focus than earlier studies, and collates the results obtained form both quantitative and qualitative sources.

Chapter 4: Research Method

Introduction

It appears from the number of research papers reviewed in the previous section that most investigations apply quantitative methods to identify unusual rater behaviour, and less attention is paid to explore the sources underlying those patterns through qualitative inquiry. Therefore, to answer the research questions, this study carried out investigations in two stages, following both qualitative and quantitative methodology. Study 1 focuses on the first set of research questions which are concerned with the psychometric qualities of the operation of the rating scale. The following questions will be discussed in Study 1.

1. Which assessment criteria generate bias of rater behaviour?
2. Which criteria elicit little variation in the distribution of the awarded scores?
3. To what extent is the halo effect or the cross-contamination of descriptor bands apparent in the distribution of scores?
4. How do raters interpret the categories of the six- point rating scale?

Study 2 investigates how raters use the rating scale by addressing the following research questions:

1. Why do assessors exhibit different rating profiles across different domains of the rating scale?
2. What construct-irrelevant factors emerge during the application of the rating scale?

The data collection and analysis will be described separately for the first study, which concentrates on the psychometric aspects of the research and is quantitative in nature, and then for the second study, which pursues a qualitative line of investigation. Thus, first the data collection for the MFR analysis will be outlined, which will be followed by the description of the collection of qualitative data in the form of unmediated concurrent verbal protocols and semi-structured interviews. In the latter part of the section, the methods applied for data analysis will be explained. The mathematical model of the MFR analysis will not be provided, however, some basic parameters which are necessary for the comprehension of the results of the data analysis will be exemplified. Appendix A also provides a glossary of the most commonly used terms in Item response theory. This section will also include the analytic framework for the analysis of the quantitative data for which sample output is presented in Appendix B.

4.1 Context of the research

4.1.1 The examination

The aim of the study was to validate the rating process and within that the analytic rating scale applied in the assessment of the subjectively scored writing task of the intermediate LSP examination administered by the Foreign Language Examination Board of the Budapest Business School. The examination was accredited in May 2000 by the Language Examination Accreditation Board and provides examinations at three levels: elementary/B1, intermediate/B2 and advanced/C1 and in seven languages: English, French, German, Italian, Japanese, Russian and Spanish. Since May 2000, more than 39,000 candidates have taken one of the language examinations with an average pass rate of 60%. Most of the test-takers are students at the Budapest Business

School. Additionally, the examinee population also includes external candidates who take the exam at one of the 18 accredited or registered examination centres belonging to the Foreign Language Examination Board. The subjects are male and female examinees between 21 and 37 years of age. They are from different parts of the country, including the capital city as well as major provincial towns and minor villages. This is the general candidate profile at the examination centre, and as the data are handled confidentially, personal details cannot be disclosed. The anonymity of the examination results was maintained throughout the research. The majority of the candidates take intermediate examinations; the most popular language is English with German to follow. The examination at each level consists of an oral and a written paper, both of which include subjectively scored tasks. The tasks are developed according to the guidelines laid down in the internal document of the Board, the Guidelines for Test Developers (Útmutató tesztfejlesztők részére, BGF, 2000), and undergo a rigorous validation process.

The Examination Board provides specific purpose language examinations in three fields: business, finance and tourism. The oral tasks are specific to each specialization, the listening and the reading parts of the written paper are common for all three specializations, and the writing tasks are in two forms: one for the business and finance examinees and one for the tourism candidates.

4.1.2 The writing paper

According to the examination specifications, the writing task may be of different text types: business letter, memo, brochure or report. The expected output of the test task is a piece of writing of maximum 200 words, written on the basis of prompts provided together with the instructions for the task. For sample tasks see Appendix C.

As with other subtests, candidates are expected to write on the answer sheet provided together with the task. Examinees are allowed to prepare a draft, but the final version should be submitted on the answer sheet, and the original task sheet with the writing prompt is destroyed after the examination for security reasons.

4.1.3 The assessment procedure

4.1.3.1 Raters

The assessment procedure is carried out according to the guidelines laid down in the Accreditation Manual of the Examination Board of the Budapest Business School (A KVIF Gazdasági Szaknyelvi Vizsgarendszerének akkreditációs anyaga, KVIF, 1999). In line with the document, the writing test papers are assessed on an analytic rating scale by two markers independently. The rating is carried out by trained raters, male and female, native speaker and non-native speaker language teachers at the Budapest Business School. They are all accredited examiners of the Examination Board, which involves regular examiner training both for the oral and the written subtest as well as standardization sessions prior to the rating. Several important considerations are taken into account in the pairing of raters. To ensure high reliability and consistency in marking, the lowest possible number of raters are applied for one type of task. However, with test papers as many as 1000 and more, this number cannot be very low. It also requires due consideration that there should be a regular rotation among raters to decrease the possibility of rater pair fossilization: a reliable within-pair marking but an across-pairs marking of questionably reliability. The members of a pair might work in complete agreement, but at the same time might be much stricter or much more lenient than another pair, also working in complete agreement. Thus, candidates might be advantaged or disadvantaged depending on the rater pair marking

their papers. In practical terms, however, in spite of sustained efforts, an ideal rotation cannot always be ensured.

4.1.3.2 The rating scale

The six-point analytic rating scale is of primary importance in terms of the current study, as the research questions focus on the rater-rating scale interaction. The assessment criteria include aspects of language use, vocabulary, discourse features and task achievement with equal weighting. The wording of the criteria is relatively straightforward, though an apparent weakness is the use of relative modifiers which might yield differing interpretations. There are six different levels within each criterion, ranging from 0 to 5. Zero score is used for no performance or for a performance which does not provide enough sample to be assessed, whereas at the other end of the scale, 5 is the maximum score that can be awarded on a performance which fully meets the requirements of the task on the given criterion. This rating scale is going to be investigated extensively in the study. The rating scale was constructed as part of the test development process prior to the accreditation of the examination. The rating scale development process lacked the state-of the art psychometric methods, for which both the theoretical background and the practical applications (Andrich, 1978a, Andrich, 1978b; Bond & Fox, 2001; McNamara, 1996; North, 2000; Wright & Masters, 1982) have become more well known and popular in recent years, especially with the appearance of the descriptor scales of the Common European Framework of Reference (Council of Europe, 2000). Due to lack of considerable previous experience and relevant expertise in quantitative rating scale construction, the descriptors were developed in an intuitive, qualitative way (Fulcher, 2003; Fulcher & Davidson, 2007) with the help of a trained expert team. More importantly, there were no useable

empirical data at the time of the construction of the scale and the accreditation of the examination system. As data collection can start only with the onset of the operation of the examination system, the real data-driven empirical validation can occur after a certain period of test administration. It is evident today from the example set by the Common European Framework of Reference descriptor scales (North, 2000) how qualitative and quantitative methods should be combined in order to create a rating scale which can be considered more valid than rating scales created by qualitative methods solely. This knowledge, however, also demonstrates the need for the validation of any operational rating scale and the rating process.

The six-point rating scale investigated in the current research includes four criteria which indirectly reflect the writing construct of the examination system and how the construct is operationalised through the tasks. The task achievement criterion includes descriptors which differentiate between levels of success attempting task completion, with a maximum score of 5 for creating a text required by the instruction, in which all points are covered logically and in full detail. At the other end of the scale in this category, zero point is awarded for a performance where no text is created at all or where the text is entirely different from what has been specified by the instructions. The second criterion is vocabulary, and the third rating category is style. For a maximum score of 5, these criteria require maximum efficiency in the use of lexis and discourse features in conformity with the expected text type. No point can be awarded on these criteria for inadequate and incomprehensible use of words, or an incoherent and fragmentary text. The final criterion concerns language use. Although the top score does not require that the text is free of mistakes altogether, a high level of grammatical accuracy and a sustained control of structures are expected. Although the tasks in the

writing paper might be of different genres across examination periods, for each task type the same criteria are used. For the complete Hungarian rating scale see Appendix D.

4.1.3.3 The rating procedure

The total score is added up from subscores awarded on the four assessment criteria. Scores awarded on this subtest range from 0 to 20. The marking session starts with the standardization procedure. Only those markers can take part in the assessment procedure who participated in the standardization session for the given task. The standardization of marking is a crucial point in the rating process. Before the standardization, the team leader with the help of a rating moderator goes through all writing performances and selects ten scripts which represent possible benchmark levels for each score. In this initial selection process, the moderators search for sample papers around the theoretical cut scores and scripts which present problems that require special attention. The standardization session itself is one-day event in which all raters marking the same type of task come together. They all individually mark the same ten scripts previously selected by the team leader and the moderator, and then, with the guidance of the team leader discuss the points awarded and come to an agreement regarding the final scores. These standardized scripts will serve as benchmark papers for the live marking and as sample performance levels for the given tasks. After the initial standardization phase of the test marking procedure has been completed, the raters work independently. The scores are registered on an assessment sheet, and after the complete test batch has been marked by the raters, test papers are reviewed again, and raters agree on the final score. In the case of differences exceeding two points, a third

assessor, an adjudicator, is requested to mark the paper. The final score is based on the two marks awarded closest to each other.

4.1.3.4 Selection of tasks and scripts for the rating procedure

Two major factors directed the selection of tasks and scripts for Study 2, one of which was field specificity. In line with their instructional practice, most markers tend to rate test papers either in business or tourism specialization. Therefore, it seemed that a task with either a less strong business focus or one in tourism specialization would be more appropriate for most of the raters involved in this project. Thus, the English task selected involved a business letter in which two potential business partners initiated cooperation. The parallel task for the German raters was a letter of application for a job (Appendix C). The other consideration in the choice of the scripts was representativeness, namely the need for the scripts to represent different levels of achievement. It was also important to bear in mind in the selection of the text that too poor a performance would result in scant data. The final set of scripts for the English and the German raters are presented in Appendix C.

4.2 Data collection

4.2.1 Data collection for Study 1, the MFR analysis

For the MFR FACETS analysis applied in the quantitative part of the study, the subscores awarded on each of the four criteria of the analytic rating scale were used. As it was confirmed in the interviews at a later stage of the data collection, in spite of the written guidelines for reaching agreement in the process of double marking, raters usually apply different strategies while agreeing on the final scores. Therefore, instead of the final agreed scores, the individually awarded scores were analyzed as these

provide a more authentic and accurate reflection of the use of the rating scale and rater behaviour.

In order to run FACETS analysis on the dataset, it was necessary to create a connected rating design (Engelhard, 1994; Engelhard & Myford, 2003; Linacre, 2003-2006; Lunz, Wright, & Linacre, 1990) which establishes direct and indirect connections between raters, eliminates subsets in the output and renders the result directly comparable. To highlight the importance of the linked design and to underline the advantages that the FACETS analysis can offer over traditional inter-rater reliability analysis, a brief comparison of the two methods, Spearman rank order correlation of classical test theory and rating reliability calculations with FACETS of modern test theory will follow.

The illustrative pairing of raters in Figure 7 indicates that interrater reliability calculations with Spearman Rho correlation do not ensure the comparison of all raters and ratings involved in the assessment procedure. Even if the markers are rotated and the fossilization of rater pairs is avoided, the results of Spearman rank correlation order analysis only provide within-pair interrater reliability indices.

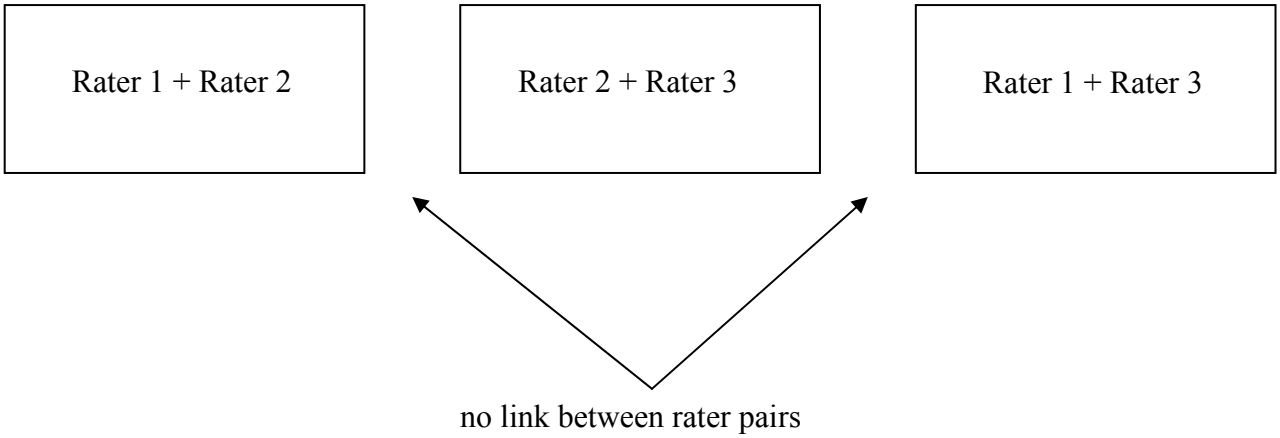


Figure 7 Lack of relationship between all raters in establishing rater reliability with Spearman's rank order correlation

It is possible to achieve relatively high interrater correlations within pairs with this approach, but there might still be significant differences between the pairs. This means that the reassuring high interrater reliability indices within pairs might mask substantial differences between pairs. In addition, this method does not yield results about rater inaccuracies, namely no data are provided on rater consistency. FACETS is capable of handling all raters and other facets in a linked design, illustrated in Figure 8.

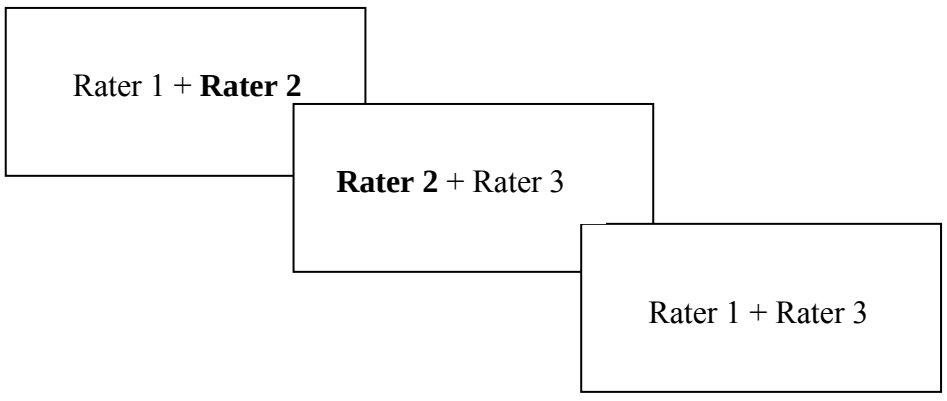


Figure 8 The connection between raters in a linked rating design

Figure 8 shows three rater pairs made up of three raters, Rater1, Rater2 and Rater3. Rater2 links Rater1 with Rater3, as indicated in the figure. Although Rater pair 3 does not include Rater2, a link has been established between all three rater pairs through Rater2 working together with both Rater1 and Rater3. Thus, this connected rating design is capable of providing input for an analysis which can compare facets of the rating process that are only indirectly linked.

As the data used in Study 1 come from the live administrations of operational tests, with a total number of 2011 scripts collected in 2004, 2005 and 2006, for practical reasons it was impossible to set up one large rating network within one period for all papers to be marked. Firstly, in order to increase the reliability of the marking, the smallest possible number of raters are used for the same task within one examination period. Secondly, writing tasks representing one level of difficulty are different along specification dimensions: candidates taking the tourism tests get a different prompt from those taking the business tests. These differences, however, are not relevant from the perspective of the rating process itself. Thirdly, only those raters take part in the rating process who participate in the standardization of the marking prior to the actual rating session. Finally, there are substantial differences in the number of papers to be marked both across examination periods and languages. All these factors contribute to the heterogeneity of the datasets used in the study presented in Table 1. What is common in the datasets is the necessary number of elements for the reliable operation of FACETS and the connected rating design. In the data selection process, I took special care to include raters in Study 1 who also participated in Study 2.

Table 1 Sources of data used in Study 1

Examination period	Language	Scripts	Markers
2004	English (tourism)	355	6
	German (tourism)	268	4
	German (business)	127	3
2005	English (tourism)	464	4
	German (tourism)	165	3
2006	English (business)	432	5
	English (business)	200	2

Altogether seven smaller linked designs were created and investigated. As the majority of the candidates take their tests in English (58 %) and German (32 %), scores collected from writing tasks in these two languages constituted the initial basic dataset. The criterion for the selection of raters was guided by the need to provide the largest possible number of scripts in a connected design. The scores awarded on each of the four criteria of the analytic rating scale were entered in a spreadsheet in the special format required by the computer program FACETS. Although during the marking procedure maximum anonymity is guaranteed for the candidates by assigning codes to the test papers, the markers names are disclosed on the marking sheets. This apparently might violate markers' personal rights, but it is essential to identify raters for practical reasons. In this way, it is possible to monitor the consistency of their work on the one hand, and on the other, this is how the connected rating design can be established. The Examination Centre assumes full responsibility that no personal details of the raters are disclosed to any uninvolved third party.

4.2.2. Data collection for Study 2

The second study sought to provide further details on rating behaviour to confirm and extend the results of the first study. Altogether 15 raters took part in this study. In the description of the study and the discussion of the results at times they will be labelled as teachers, but besides being language teachers, they are all accredited language examiners of the Foreign Language Examination Centre of the Budapest Business School. With one exception, they are all female examiners who have been acting as test developers, markers and interlocutors in the examination system since the accreditation of the exam in May 2000. They represent the core examiner cohort, and two of them are chief examiners. Seven of them pursue PhD studies, three of them in language testing. The over-qualification of the sample might question the generalizability of the study, yet they were selected as research participants for two reasons: firstly, because they are the examiners whose involvement in every aspect of the work of the examination centre has been the most direct and intense for the past years, and secondly, because with their participation, the dropout rate in this part of the study could be reduced to practically zero. In addition, it was expected that due to their close involvement and genuine interest in testing issues, they would be more willing to adopt an honest approach in voicing negative feelings and controversial attitudes to the existing rating practice. Their heightened awareness and professional attitudes to the questions raised, however, will be considered as a possible intervening factor in the interpretation of the results. Before the onset of the data collection, an informed consent of the participants was sought. They were briefed about the purpose of the study and the procedures applied, informed about the prospective benefits and applicable practical results, and were also ensured of maximum confidentiality in handling their data. Four of the participants in Study 2 were teachers of German, and 11 markers were teachers of

English. Although the inclusion of teachers of German slightly decreased the number of comparable results in this study, but it was felt that as Study 1 also investigated the rating profile of teachers of English and German, the same procedure should be followed in Study 2 as well.

Data on rating behaviour were obtained in three different ways. First, 15 raters were asked to rate three writing papers while verbalizing their thoughts. With a full awareness of the possible shortcomings (Ericsson & Simon, 1984) and criticism levelled against the think aloud method (Nisbett & Wilson, 1977; Green, 1998), this procedure was expected to reveal details of rating behaviour which could be triangulated with data obtained from other sources. Each rater was given a full demonstration of the think aloud procedure prior to marking. Then, they were given the rating pack including the three scripts to be marked, the original task, the sample solution, the rating scale and the marking sheet. They were also equipped with a digital voice recorder. The concurrent think-aloud protocols were carried out in an unmediated way (Green, 1998), and were recorded on the digital recorders. The Voice Operated Recording (VOR) facility, which stops the recording automatically during silent pauses, was turned off to ensure that the complete process is recorded and so the marking times can also be compared. As the rate of speech also varies rater by rater, Table 2 provides details about the number of words of the transcribed protocols. The raters marked the same three scripts, yet it seems that the time of rating shows considerable differences: the fastest rater completed the task in 7 minutes and 25 seconds, whereas the longest time needed for the marking was 1 hour 32 minutes and 43 seconds.

Table 2 Details of the think aloud data in Study 2

Rater identification	Approx. number of words	Time needed
Rater 1	1023	25’ 25”
Rater 2	3464	38’ 25”
Rater 3	1216	13’ 44”
Rater 4	1232	19’ 19”
Rater 5	2305	26’ 43”
Rater 6	1184	16’ 6”
Rater 7	10830	1 hour 32’ 43 “
Rater 8	672	7’ 25”
Rater 9	1485	25’ 20”
Rater 10	1053	19’ 2”
Rater 11	1288	21’ 30”
Rater 12	575	19’ 19”
Rater 13	1524	16’ 56”
Rater 14	2097	22’ 51”
Rater 15	1628	25’ 19”

The second set of data to be investigated came from the marking sheets, on which raters recorded the scores they gave for the performance on the writing task. Here we had comparable data for the same three English letters (Appendix E) marked by eleven teachers of English and the same three German scripts (Appendix F) marked by four teachers of German.

Thirdly, after the marking procedure the raters were invited for an interview in which they were questioned about their perceived rating behaviour. The semi-structured interviews focused on participants’ general rater behaviour rather than how they responded to the particular tasks they had to mark. The interview protocol consisted of two subsections: the first set of questions was based on aspects of rater misbehaviour identified by Linacre (2003-6). Respondents were asked how they assessed their own

level of severity. An attempt was made to identify the factors which elicited unusual rating behaviour from the markers: details in the written performances that pushed them into one of the extremes, or left them undecided and encouraged them to stay in the safe mid-range of scores. The implied aim of asking raters to put the assessment criteria in the order of importance was to spot the knock-off or the halo effect of one dominant criterion. In the second part of the interview, raters were asked to assess the usefulness and the ease of use of the criteria applied in the assessment of the writing task. At the end of the interview, respondents were invited to comment on any relevant aspect of the subjective rating process that they felt had not been included in the interview. For details of the interview protocol, see Appendix G. The interviews were conducted in Hungarian, and although the same questions were put to each interviewee, the length of the interviews varied between 11 minutes 54 seconds and 42 minutes 58 seconds. Table 3 presents more details about the interview data. The interviews were all digitally recorded, and the interviews were fully transcribed.

Table 3 Details of the interview data for Study 2

Rater identification	Approx. number of words	Time needed
Rater 1	3392	28’ 20”
Rater 2	6255	42’ 12”
Rater 3	1391	11’ 54”
Rater 4	5795	42’ 58”
Rater 5	5138	31’ 50”
Rater 6	3916	26’ 9”
Rater 7	3360	23’ 9”
Rater 8	1713	12’ 19”
Rater 9	2994	25’ 33”
Rater 10	3367	22’ 41”
Rater 11	1871	13’ 47”
Rater 12	3406	26’ 2”
Rater 13	3108	22’ 21”
Rater 14	6785	26’ 21”
Rater 15	2585	17’ 9”

4.3 Data analyses

4.3.1 Data analysis for Study 1

Many-faceted Rasch analysis was carried out with the help of FACETS (Version 3.61.0) software in the first study. The basis of MFR analysis is the assumption that the score obtained on a subjectively rated performance is influenced by several factors. The factors or facets that might affect the final score include the ability of the student, task difficulty or rater harshness and various other aspects of the test. Many-faceted Rasch measurement is capable of taking into account all of these factors in an attempt to arrive at a score that best approximates the candidate’s performance. In the validation process it is possible to investigate the rater or the rating performance, the candidate, the task or

the rating scale. For the analysis in Study 1, scores on the six-point analytic rating scale with four criteria were used. In the data analysis process, FACETS makes a number of iterations through the dataset, which is the set of analytic scores awarded on a performance by two different raters on four assessment criteria. During each iterative phase, the program attempts to bring observed values and expected values, that is, the true score based on probability, together as much as possible. In each iterative phase the residual between the observed score and the expected score decreases. If this is successful, convergence is reached; if not, there is likely to be a large amount of unexplainable variance. This computer program attempts to identify facets or aspects of the dataset that do not fit the measurement model statement, hence the name of the programme. Even if convergence is reached, the dataset might include misfitting elements which are identified in the output files. Since the current validation concerns rater and rating scale interaction, the facets under investigation were the raters and the different assessment criteria. As it was formulated in the research questions, certain criteria might be used more extensively than others, and certain criteria might not prove to be genuinely useful in assessing the given task. Special attention is attributed to the hypothesized existence of the possible dominance of one assessment criterion, or in other words, the halo effect.

In the analysis, the model statement which delineates the direction of the investigation and establishes the basis of the probabilistic model to which the data is expected to conform is as follows:

Models=?,#,?B, R6

The first question mark signifies the candidate, the hash mark stands for the rater, the second question mark stands for the assessment criteria with letter B suggesting that

bias is to be sought during the analysis, and the final R6 letter and number combination defines that the scores vary in six different forms, namely on a scale ranging from 0 to 5. A line from the dataset consistent with the above model statement looks like the following string:

1, 2, 1-4, 2,3,3,4

The above line reads as follows: Candidate 1 rated by Rater 2 assessed on four criteria received 2, 3, 3 and 4 points on those four criteria respectively. The special character # for the rater facets allows for a detailed investigation of the "personal understanding of the rating scale" (Linacre, 2003-6), and unusual rating patterns in the use of the rating scale can be identified, such as frequent or infrequent use of categories or non-sequential interpretation of the categories.

The following questions were addressed to investigate rater and rating scale interaction in Study 1.

1. Which assessment criteria generate bias of rater behaviour?

In the bias analysis to locate discrepancies in the rating pattern absolute standardized z or t scores greater than 2 indicate significant rater and criterion interaction effect. It is important to note, however, that only regular occurrences of the same rater criterion effect should be considered bias.

2. Which criteria elicit little variation in the distribution of the awarded scores? In order to provide an answer to this question a thorough analysis of fit statistics is required: low

infit, or overfit indicates malfunctioning category in the rating scale domain by suggesting lack of variability in the scores given on the criterion.

3. To what extent is the halo effect or the cross-contamination of descriptor bands apparent in the distribution of scores?

As an extension of the previous question, a further analysis of fit statistics might reveal overfit which indicates little variation in the scores: a small range of scores across a candidate or clustered scores for certain criteria indicate the halo effect.

4. How do raters interpret the categories of the six- point rating scale?

The analysis of category fit and the graphic representation of the probability curves for the six scale steps reveal raters' personal interpretation of the scale and the possible misinterpretations of certain categories.

The results of the analysis of these four aspects of the rater rating scale interaction will be discussed in turn in the Results and Discussion chapter.

4.3.2 Data analysis for Study 2

In the second study rater and rating scale interaction was investigated on the basis of data obtained from three different sources: scores awarded on writing performances, think-aloud data collected during the rating process and interview data related to the raters' perceived rating behaviour. Although occurrences of rater misbehaviour had been identified in Study 1, these results only delineated a general interaction profile which highlighted the major focal points to be considered in the next stage of the research. The majority of the data in the second study are qualitative in

nature and are of two different types. The interview data are researcher-provoked data as the process of the "interviews involves actively creating data which would not exist apart from the researcher's intervention" (Silverman, 2001, p. 159) as opposed to the data of the think aloud protocols, which yielded naturally occurring data.

The qualitative data were analyzed according to the steps suggested by Miles and Huberman (1994). The initial data reduction was followed by the display of the data in the form of arranging the relevant coded material in a matrix. The cycle of data analysis was completed by drawing conclusions with regard to the research questions raised in Study 2. For the qualitative analysis the computer program Maxqda2 (2005) was used. With the help of this program, it was possible to assign selected text segments to the predefined codes created by the researcher. It also established links between the codes on the basis of the frequency of co-occurrences, besides summarizing category frequencies and tallying word frequencies. Samples from the qualitative analyses are presented in Appendix H. From the large number of functions this software is capable of performing, however, only a limited number of analyses were used. Both the think aloud protocol transcripts and the interview transcripts were imported into the programme and fully coded. In order to render the coding more accurate and reliable, while I was looking at the scripts the recording was played on the computer simultaneously. This double-channelled presentation of the source text during coding assisted in resolving ambiguities in the written texts. In addition, closeness to the data, the lack of which is a frequently voiced criticism against qualitative researchers using software for data analysis (Weitzman, 2000) was thus facilitated.

4.3.2.1 Think-aloud data analysis

Data analysis for the think aloud protocols proceeded along the lines described above. The initial step in the analysis was the establishment of the categories. The development of the coding scheme was an iterative process. Having established the main categories as suggested by the results of Study 1 and the theoretical framework of the research, the partitioning of variables (Miles & Huberman, 1994) seemed necessary. With a full awareness of the coding schemes applied in similar studies (e.g. Cumming, 1990; Cumming, Kantor, & Powers, 2001; Erdosy, 2004; Lumley, 2005; Vaughan, 1991), a new coding scheme was created. Although the data could have yielded far more categories than the final scheme, the chief guiding principle in setting up the categories was a clear focus on the research questions. The categories were intended to focus on rater and rating scale interaction during the assessment procedure. Table 4 presents the coding scheme for the think aloud data. The four major categories include further subcategories which are presented in the third column. The final column provides explanation for those categories of which the labels do not give clear guidance.

Table 4 Coding scheme for the TA transcripts

Code	Category	Subcategory	Reference
1.	1.Performance dimension		
1.1		**Candidate**	
1.1.1		Hypothesizing about	Making guesses about general candidate profile
1.1.2		Advice/teaching	Commenting on mistakes by giving advice on what should have been done
1.2		**Criteria**	The four assessment criteria in the rating scale
1.2.1		Task achievement	Reference to task achievement
1.2.2		Vocabulary	Reference to vocabulary
1.2.3		Style	Reference to style
1.2.4		Language use	Reference to language use
1.2.5		CEF levels	Reference to the scale descriptors of

		the Common European Framework of Reference
1.2.6	Problems with	Difficulties encountered during the application of the criteria
1.3	**Task/prompt**	
1.3.1	Relevance	Reference to task relevance
1.3.2	Instruction	Reference to and problems with the instructions accompanying the task
1.3.3	Problems with	Difficulties with the original task
1.4	**Performance/text**	
1.4.1	Length	Reference to the number of words
1.4.2	Layout	Comments on text format
1.4.3	Lack of/presence of	Inclusion or omission of relevant information as described by the prompt
1.4.4	Lifting	Including parts of the prompt literally in the text
1.4.5	Specific comments	Detailed explanation, explicitation
1.4.6	Overall impression	General opinion
1.4.7	Reading or reference to the actual text	The actual reading of the text
1.4.8	Positive remarks	Expressing satisfaction
1.4.9	Negative remarks	Expressing dissatisfaction
1.4.10	Neutral commentary	Explanation
1.4.11	Prefabricated memorized chunks	Commenting on verbatim repetition of the prompt
1.5	**Rater**	
1.5.1	Self	Comments on own rating
1.5.2	Pair	Reference to the rating pair
1.5.3	Rater group/standardization	Reference to the standardization process preceding the actual marking
1.6	**Rating process**	
1.6.1	Sequence	Steps in the rating process
1.6.2	Rating process technicalities	General procedures applied in the rating sequence
1.6.3	Rhetorical question	Questions expressing emotions, incomprehension, surprise
1.6.4	Tech talk	Reference to rating technicalities
1.6.5	Comparison	Relating the rated performance to other performances
1.7	**Score**	
1.7.1	Analytic	Awarding scores according to the analytic criteria
1.7.2	Total	Reference to total score
1.7.3	Pass mark	Mention of the pass mark
	Problems with	Difficulties encountered while giving scores
1.7.4	Assessment sheet	Reference to the use of the

			assessment sheet on which the analytic scores are registered
1.7.5		Pass mark	Mention of the cut off score
	2. Rating-irrelevant comments		
2.1		External circumstances	Reference to the environment
2.2		Emotionalism	Critical remarks, attitudes
2.3		Recording the TA process	Reference to the voice recording procedure
2.4		Complete diversion	Distractions
2.5		Fillers, hesitations	Verbalized time gaining devices
2.6		Reference to this particular rating occasion	Commenting on actual rating process
3		Indecipherable	Incomprehensible chunks in terms of the categories
4		Explaining rater behaviour	Justifying actions

As it is apparent from the table, four major categories with subcategories were set up at the outset of the analysis. The sub-categories included the elements of the writing performance dimensions (1), aspects of rater behaviour (4), rating-irrelevant features (2), and indecipherable text segments (3). The final set of codes emerged as an attempt at a theory-driven analytical framework. Aspects of rater misbehaviour and the empirical results of Study 1 were meant to be incorporated in the coding scheme, yet it turned out that this would lead to an information overload that is rather difficult to manage. These aspects of the investigation were withheld for the interview data analysis.

Consistent with Lumley's view (2005) in defining the units of analysis, it seemed that for the purpose of the analysis in this study, a pragmatic rather than a linguistic approach should be adopted. Thus, the segmentation was primarily content driven rather than linguistic unit based. In other words, the text itself was not initially unitized, but the free-flowing script was segmented along the category divisions. As a

result, the segments to be coded were of different length, “chunks of varying size – words, phrases, sentences, or whole paragraphs, connected or unconnected to a specific setting (Miles & Huberman, 1994, p.56)”. In order to obtain meaningful units, at times complete question and answer turns, short though they might have been, were coded as one segment.

With the exclusion of the more subjective aspects of the rating process which might have called for extensive hypothesizing and inferring rather than categorizing on descriptive grounds, the need for double coding seemed to be less compelling. Instead of applying a second coder in the analysis, a rigorous single-coding procedure was worked out, and reliability was established by the coding-recoding method for the initial stage of the analysis. Three texts were selected to establish the validity of the categories. Although the think aloud protocols resulted in ample data, they were rather different in terms of usability. The fifteen accounts fell into two broad categories: the minority of them might be labelled as “staged rating” with traces of the Hawthorne effect and an attempt to fulfil researcher expectancy. The majority of them, however, provided genuine “stream-of-consciousness disclosure of thought processes” (Milanovic, Saville, & Shuhong, 1996) type of data which proved to be a rich source of information for the study. Three of these rich texts were selected and coded twice with a one-day time span, and then the two versions were compared for discrepancies. This form of establishing intra-coder consistency also contributed to the refinement of the final categories of the coding scheme. At this stage of the instrument validation, the coding was carried out manually, and on a paper and pencil basis. Doing analysis by hand, in Weitzman’s view (2000) is a good initiation into a more sophisticated mode of qualitative data analysis. Following this initial stage, the complete material was coded. The coding procedure was carried out with the computer program Maxqda2 (Version

2.0). The analysis itself allowed for parallel procedures: besides categorizing and coding the text segments, certain segments were colour coded and annotated with memos. The memo function of the program, in line with Glaser's concept of theoretical memoing (1998), allows the researcher to generate ideas and record them in the form of memos during the data collection and analysis procedures. Thus the memos constitute a collection of ideas which help to connect ideas, establish relationships when interpreting the data and drawing conclusions. Instead of mechanically rendering text segments to categories, the coding process itself was data analysis, as Miles and Huberman (1994) also claim. The data from different sources, namely the highlighted text chunks, the thematically coded segments alongside with the memos helped to establish links in the data within the study and across the two studies.

4.3.2.2 Interview data analysis

The procedures applied in the analysis of the interview data were highly similar to those in the think aloud data analysis. The semi structured interviews also supplied immensely rich data for analysis. The semi structured interview format allowed for the majority of the conversations to assume a personal tone, which highlighted unexpected aspects of the research questions. The fully transcribed interview data were also analyzed with the Maxqda2 software. The creation of the categories was largely driven by the aspects of rating behaviour which had been omitted from the analysis of the think aloud data. As the interview questions were explicitly directed towards the identification of rater misbehaviour and raters' perception of their own rating practices, the first eight categories were related to the interview questions. In comparison with the coding system for the think aloud data, the categories used in this analysis were more directly concerned with rater characteristics.

The first major category, leniency and severity (1), includes five subcategories. Apart from general comments on leniency and generosity (1.1), in the process of double scoring, a comparison with the rater pair (1.2) warrants due attention. The same issue is further elaborated and investigated by reference to the agreement procedure (1.3) whereby raters agree on the final scores to be awarded for a piece of writing. As rating is carried out on an analytical scale by both readers, their attitude to the analytical rating scale and global assessment (1.4), and how these two different issues feature the rating process are also worth considering. The final category discussing rater leniency and harshness focuses on possible assimilation to the rater pair (1.5), and the reasons and the consequences of such a strategy.

The second major group of categories is concerned with extremism and central tendency. Besides talking about zero and maximum score (2.1), the reasons for giving such scores are identified (2.2). The frequency of these scores in the interviewee's rating history is also investigated (2.3). The halo effect (3), or the contamination of descriptor bands is highlighted by identifying the most important criterion (3.1) and the least important criterion (3.2) in the rating scale. A more indirect link was expected to be established through the discussion of the possible relationship between criteria (3.3) and what impact they might have on one another. Response sets (4), and the playing it safe strategy (5), when raters show unusual closeness to their pair's scores constitute further categories in the analysis of rater misbehaviour. The instability code group (6) embraces various factors influencing assessment (6.1), those resulting in unusual strictness (6.2) and leniency (6.3). Two different aspects of marking consistencies are examined: within rating period consistency (6.4), which includes features that might influence the stability of rating within a short timeframe, and consistency across rating periods indicating fluctuations over a longer period of time (6.5). A further category

collects evidence for the blackout syndrome (6.6). This subcategory includes references to the existence and the possible sources of unusually large discrepancies between the raters observed in the case of a limited set of consecutive papers within a larger batch of papers. Specifying the usefulness (7) and the ease of application (8) of the assessment criteria is highly relevant from the perspective of the focus of the research on the one hand, and it also helps cross-validate similar data both in the interviews and in the think aloud data, on the other. The most technical of all of the codes is the technicalities (9) category, which accumulates text segments describing the process of recording the interview, manipulating the recorder, showing the document under discussion to the interviewee. Diversions (10) as a separate category, was set up in order to gather all remarks showing no connection with the research questions. Unlike the previous category, emotionalisms (11), a slightly arbitrarily coined word reflects emotional outburst or manifestation generated by some aspect of the rating process. Four categories are directly related to the data collection method, the interview. Category 12 embraces the interview questions, both those included in the interview schedule (12.1) and those non-scheduled questions (12.2) which arouse during the interview. Whereas the interviewees sought for clarification (13.1), the interviewer offered clarification, or got engaged in an explanation or exemplification of a certain question or statement (14.1). The questions, which were either repeatedly put to the interviewee or asked in a more provocative way to ensure a direct answer, fall in a separate group (14.2). The final four categories are all directly linked to the rating process. The use of the Common European Framework of Reference (15) and its implications were also highlighted by some of the raters. The final three categories concern different elements of the marking procedure. The rating sequence (16) sheds light on what steps markers take during the assessment procedure, whereas the rating process and scale category (17) includes all

those statements which cannot be classified in any other category. In spite of the criterion referenced nature of the examination, interviewees repeatedly made comparisons between performances (18). Table 5 summarizes the categories applied in the analysis of the interviews.

Table 5 Coding scheme for the interviews

Category
1 Leniency, severity
1.1 Leniency, generosity
1.2 Comparison with rater pair
1.3 Agreement procedure
1.4 Global and analytic rating
1.5 Adjusting to rater pair
2 Extremism, central tendency
2.1 Zero/maximum score
2.2 Reasons for zero/maximum
2.2.1 Zero
2.2.2. Maximum
2.3 More frequent
3 Halo effect
3.1 Most important criterion
3.2 Least important criterion
3.3 Relationship between criteria
4 Response sets
5 Playing it safe
6 Instability
6.1 Factors influencing assessment
6.2 Factors resulting in strictness
6.3 Factors resulting in leniency
6.4 Within rating period consistency
6.5 Across rating period consistency
6.6 Blackout
7 Usefulness
8 Ease of application
9 Technicalities
10 Diversion
11 Emotionalism
12 Interview questions
12.1 Scheduled interview questions
12.2 Non-scheduled interview questions
13 Extra comments made by the interviewee
13.1 Asking for clarification
14 Extra comments made by the interviewer
14.1 Clarifying question/statement
14.2 Provoking questions/statements
15 CEF
16 Rating sequence
17 Rating process/ scale in general
18 Comparison

As it has been stated before, more categories could have been established and the qualitative data could have been investigated from more aspects, but the current analysis was restricted to the main focus of the study. It cannot be left unsaid at this point, however, that the material offers further information and rich details on rater behaviour for scrutiny at a later period and for a different line of investigation.

4.4 Conclusion

The rendering of data along the lines outlined above was expected to offer a convincing answer to the second major research question which sought to provide information about rater behaviour, and explore how different assessment criteria are interpreted and applied by the raters. The data ordered according to these principles and arranged in the predefined categories were then collated with the results of the quantitative investigations of Study 1, and this triangulation of results was carried out to allow drawing conclusions based on the research questions.

Chapter 5: A Quantitative Approach to Rater Misbehaviour

Introduction

This chapter presents the results of Study 1, the Rasch-based investigation of the quantitative data of the writing test scores. The results of 2011 writing test performances were analysed with the help of MFRM in Study 1. The FACETS analysis provided numerical data about rater misbehaviour and about the functioning of the assessment scale. Raters' leniency and harshness was established, as well as their marking consistency. The rating scale analysis also provided insight into the functioning of the assessment instrument. All four criteria, task achievement, vocabulary, style and language use, as well as the six score categories, 0-5, were examined. The fit statistics showed the amount of distortion in the measurement system, in other words the difference between the observed values and the model expectation. In order to render the results more comprehensible, a brief explanation of some of the FACETS output parameters will be given with special emphasis on rater misbehaviour and on rating scale use. At the end of the chapter, the research questions related to the validity of the rating scale will be answered.

5.1 FACETS

The present study confines itself to a principally tailored use of the analyses that the program can offer. In order to identify rater misbehaviour as suggested by the research questions, only a limited set of parameters will be investigated. The appropriate functioning of the raters and the appropriate functioning of the rating scale, within that the steps (zero to five) and categories (assessment criteria) were examined. Appendix B includes excerpts from an original output file for one of the analyses

carried out. The description of the most important elements of the output will follow, but the explanation will be confined to those statistics only which are directly related to the research questions. For purposes of giving a brief explanation of the FACETS output, the original format of the tables is retained, but in the second part of the chapter the tables are simplified. To ensure anonymity for the participants, raters were coded. The names of those participating in both studies were listed in alphabetical order and assigned a number (for example Rater2), those only taking part in Study 1 were coded according to a different principle. Their names were also arranged in alphabetical order, and they were numbered, but to differentiate them from the first cohort, a letter in their codes also indicated what language they taught. Thus, RaterG3 is a teacher of German who only participated in the linked design of the first study. This explains the different system of codes in the tables.

The first two tables in the Output Description in Appendix B provide basic information related to the analysis, which include such technical details as the number of elements in the analysis and the names and location of various files used. The third table is the first meaningful one from the point of view of the analysis: the two types of iterations are listed which end in the data and the model reaching convergence. The process of fitting the data to the model means reducing the residuals between the observed and the expected values, which is manifested in the decreasing numbers in the table. The iterations stops when the residual cannot be further decreased in a meaningful way. The number of iterations can be defined in advance in the specification file. Either a certain number is set for the iterations or the program is allowed to carry out the number of iterations necessary for reaching maximum convergence. As it is apparent in this file, for the present analysis an unlimited number of iterations were set. This resulted in 389 iterations, which is not an unusually high

number with regard to the amount of data. This analysis was completed in 36 minutes and 53 seconds. Two types of iterations are carried out in the analysis: initially approximate (PROX) estimates are made to obtain rough estimates which are followed by joint or unconditional maximum likelihood estimates (JMLE) to get more accurate estimates. This table also gives information of the subset connection (Subset connection O.K.), namely that the data can be interpreted within one frame of reference. This is one of the most important aspects of the analysis, as only a perfect subset connection can ensure direct comparability of the data. Where there is no complete linkage between the raters in a rating situation, it is possible to reach convergence, but there will be different subsets only weakly linked to each other. Alternatively, it is also possible to impose artificial links on the dataset. In all the analyses run in this project, perfect subset connections were established for the data. The data presented in Table 6 are central to the research questions and will be discussed and interpreted in more detail. This "All facets vertical rulers" output provides a global visual representation of all facets in the analysis.

Table 6 All facets vertical ruler from the FACETS output

```
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1*,2A,3A) Yardstick (columns,lines,low,high)= 0,2,-8,10
-------------------------------------------------------------------------------------------
|Measr|+Student  |-Rater  |-Ctiteria                                | S.1 | S.2 | S.3 | S.4 |
-------------------------------------------------------------------------------------------
+  10 +          +        +                                        + (5) + (5) + (5) + (5) +
|     | .        |        |                                        |     |     |     |     |
+   9 +          +        +                                        +     +     + --- +     +
|     | *        |        |                                        |     | --- |     |     |
+   8 +          +        +                                        +     +     +     + --- +
|     | .        |        |                                        |     |     |     |     |
+   7 + **.      +        +                                        +     +     +  4  +     +
|     | *.       |        |                                        | --- |     |     |     |
+   6 + *.       +        +                                        +     +  4  +     +  4  +
|     | **.      |        |                                        |     |     |     |     |
+   5 + **       +        +                                        +     +     + --- +     +
|     | *        |        |                                        |  4  |     |     |     |
+   4 + ***.     +        +                                        +     +     +     + --- +
|     | **       |        |                                        |     | --- |     |     |
+   3 + ****.    +        +                                        +     +     +     +     +
|     | *****    |        |                                        | --- |     |     |     |
+   2 + ****.    +        +                                        +     +     +  3  +  3  +
|     | ***.     |        |                                        |     |  3  |     |     |
+   1 + ******.  + Rater13+                                        +  3  +     +     +     +
|     | *****    | RaterG4| Style                                  |     |     |     |     |
*   0 * *******. *        * Grammar          Task achievement     *     *     *     * --- *
|     | *****.   | Rater7 | Vocabulary                             | --- |     |     |     |
+  -1 + ******.  + RaterG2 +                                       +     + --- + --- +     +
|     | ****.    |        |                                        |  2  |     |     |     |
+  -2 + ***.     +        +                                        +     +     +     +     +
|     | ****.    |        |                                        |     |  2  |  2  |  2  |
+  -3 + *.       +        +                                        + --- +     +     +     +
|     | **.      |        |                                        |     |     |     |     |
+  -4 + .        +        +                                        +  1  + --- +     +     +
|     | *.       |        |                                        |     |     | --- |     |
+  -5 + .        +        +                                        +     +     +     + --- +
|     | *        |        |                                        | --- |  1  |     |     |
+  -6 + .        +        +                                        +     +     +     +  1  +
|     |          |        |                                        |     |     |  1  |     |
+  -7 + .        +        +                                        +     + --- +     +     +
|     | .        |        |                                        |     |     |     | --- |
+  -8 +          +        +                                        + (0) + (0) + (0) + (0) +
-------------------------------------------------------------------------------------------
|Measr| * = 3    |-Rater  |-Ctiteria                                | S.1 | S.2 | S.3 | S.4 |
-------------------------------------------------------------------------------------------
```

The first column represents the common yardstick of measurement, the logit values. Conventionally, a mean of 0.0 is set as the mean item difficulty, and the top end of the scale represents more endorsement, such as more able student, more difficult item and stricter rater. The lower end of the scale consequently locates less able candidates, less harsh raters and less difficult items. The final column displays the scores in terms of the rating criteria. A quick scan of this graphic representation of the

data leads to the following conclusions: student ability distribution approximates the Gaussian curve. Some differences can be observed between the raters in terms of leniency, but these differences are not substantial. Rater13 is the strictest, followed by RaterG4. Rater7 and RaterG2 are apparently less harsh. The rating criteria are neatly clustered around the mean, yet there is a minor difference between how easily candidates can get a high mark on each of them. It seems from the table that raters were the harshest on the style criterion, and it was the easiest to obtain high scores on the vocabulary criterion. Grammar and task achievement are at the same level of difficulty around the mean. In the subsequent part of the chapter, it will be examined whether this is a general pattern in the case of the rating criteria, or whether there is a fluctuation in raters' use of the assessment criteria. Further subtables provide more detailed information about the different facets in the analysis. Central to the investigation, however, is the next table (Table 7), which shows how consistent raters are and how harsh an attitude they adopt in their rating practice.

Table 7 Rater Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit		Outfit		Estim. Discrim	Exact Agree.		N Rater
						MnSq	ZStd	MnSq	ZStd		Obs %	Exp %	
1613	532	3.0	2.97	-.94	.09	.85	-2.6	.82	-2.8	1.16	68.4	56.1	4 Rater2G
1179	400	2.9	2.96	-.73	.11	1.01	.1	.99	-.1	.98	72.8	58.0	3 Rater7
1368	536	2.6	2.72	.74	.07	1.03	.5	1.03	.5	.97	50.7	48.0	1 RaterG4
1734	668	2.6	2.75	.93	.07	1.08	1.4	1.09	1.5	.91	51.6	48.5	2 Rater13
1473.5	534.0	2.8	2.85	.00	.09	.99	-.1	.98	-.2			Mean (Count: 4)	
215.2	94.8	.2	.12	.84	.01	.09	1.5	.10	1.6			S.D. (Populn)	
248.4	109.4	.2	.13	.97	.02	.10	1.8	.12	1.9			S.D. (Sample)	

Model, Populn: RMSE .09 Adj (True) S.D. .84 Separation 9.58 Reliability (not inter-rater) .99
Model, Sample: RMSE .09 Adj (True) S.D. .97 Separation 11.07 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 386.8 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 3.0 d.f.: 2 significance (probability): .22
Inter-Rater agreement opportunities: 1068 Exact agreements: 636 = 59.6% Expected: 555.7 = 52.0%

The first numeric column in the chart (Observed Score), which gives the total number of scores awarded by the rater, is followed by the total count (Observed Count), which

is the total number of papers marked. The third column provides the observed average score on one criterion given by the marker. The same data are presented in the following column: the adjusted score including the measurement error. The column labelled 'Measure' gives the data on the logit scale, followed by the standard error. In my example the leniency of the raters ranges between -.94 and .93 logits, with Rater 2G the most lenient and Rater13 the strictest.

Fit statistics are central to the interpretation of the data. They are Chi-square statistics, which describe the fit of the data to the model, and represent the residual between the observed and the expected counts. A technical yet comprehensible explanation of the calculation of fit statistics is presented in Bond and Fox (2001). Outfit and infit statistics are the commonly used labels for the Chi-square statistics in IRT. Outfit is calculated as the sum of squared standardized residuals and includes outliers. Infit is the information weighted sum and is sensitive to inliers. Therefore, infit is the commonly reported statistics and should be considered as the basis of data interpretations. The fit statistics are presented in two different forms: as MnSq (mean square) and ZStd (Z standard deviation). "Mean-squares show the size of randomness, i.e. the amount of distortion of the measurement system, Zstd are t-tests of the hypothesis 'do the data fit the model (perfectly)?'. They are reported as z-scores, i.e., unit normal deviates. They show the improbability (significance). 0.0 are their expected values. … If mean squares are acceptable, them Zstd can be ignored" (Linacre, 2006, p. 127). Infit statistics expectation is 1, range 0 to +infinity. Less than 1, overfit, indicates too little variation, in other words lack of independence. Over 1, misfit, represents excess variation - this becomes critical when it reaches Mean +(2 x SD). In order to identify the misfitting raters or criteria, the general rule is to consider those over 1.3 misfitting. Overfit results in figures below .7. These are the most commonly accepted

boundaries in language testing research (McNamara, 1996), but variations for the acceptable range of fit exist. Linacre (2003-6) and Weigle (1998) define the range for items falling between .5 and 1.5 as productive for measurement. Items below 0.5 are less productive but not degrading, those between 1.5 and 2 are unproductive but not degrading and items beyond 2 are distorting for measurement (Linacre, 2003-6). The acceptable range in fact is sample and situation dependent, and for sample sizes exceeding 30 it is between the mean ± twice the standard deviation of the mean square statistic. Wright and Linacre's (1994) suggestion concerning the acceptable ranges for fit statistic in different situations is presented in Table 8.

Table 8 List of acceptable fit statistic for different instruments

Reasonable Item Mean-square Ranges for INFIT and OUTFIT	
Type of Test	Range
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

As the sample sizes in my study are of varying size, the general range between .5 and 1.5 will be considered acceptable. In sum, misfit (mnsq> 1.5) suggests lack of rater consistency, and overfit (mnsq < 0.5) indicates lack of variability, that is, less variability in the data than the model predicts. Misfitting items are also labelled as noisy, and overfitting elements are often called muted. The infit statistics all fall within the accepted range, no inconsistencies can be detected in this dataset among the raters. FACETS does not calculate traditional inter-rater reliability if not specifically requested

in the specification file but shows the extent to which the expected agreement between raters differed from the observed agreement. Linacre (2003-6) treats too high an agreement between raters as a negative feature, as this indicates lack of independence and the fact that the raters are acting as scoring machines. It may be concluded that a desired and acceptable level for rater agreement should equal or exceed the exact agreement level but be lower than all agreement opportunities. In the sample dataset the observed agreement percentage slightly exceeds the level of the expected agreement. This agreement according to Linacre (2003-6) is similar to Cohen's Kappa agreement index. Rasch Kappa is calculated in the following way:

Rasch Kappa = (observed agreement% - expected agreement %) / (100 - expected agreement %)

If the data fit the model perfectly, the expected agreement corresponds to the model expectations and the Rasch Kappa value is 0. Negative Rasch Kappa values indicate a degree of agreement which is much less than the model expectation, and agreement indices much higher than 0 imply an unnecessarily high agreement. In the example above the Kappa Rasch agreement index is .15, close enough to 0, the model expectation. Table 9 represents the criteria measurement report, and displays information about the functioning of the individual criteria.

Table 9 Sample criteria measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	N Criteria
1549	534	2.9	3.01	-.53	.08	.75	-4.3	.73	-4.4	1.25	2 Vocabulary
1489	534	2.8	2.92	-.11	.08	1.18	2.7	1.16	2.2	.83	4 Grammar
1460	534	2.7	2.87	.10	.08	1.11	1.7	1.11	1.5	.89	1 Task achievement
1396	534	2.6	2.77	.54	.08	.96	-.6	.96	-.5	1.03	3 Style
1473.5	534.0	2.8	2.89	.00	.08	1.00	-.1	.99	-.3		Mean (Count: 4)
55.1	.0	.1	.09	.38	.00	.16	2.7	.17	2.6		S.D. (Populn)
63.6	.0	.1	.10	.44	.00	.19	3.1	.19	3.0		S.D. (Sample)

Model, Populn: RMSE .08 Adj (True) S.D. .37 Separation 4.47 Reliability .95
Model, Sample: RMSE .08 Adj (True) S.D. .43 Separation 5.20 Reliability .96
Model, Fixed (all same) chi-square: 84.0 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 2.9 d.f.: 2 significance (probability): .23

The data shown in the first graphic FACETS output (Table 6) is spelled out in full in Table 9. The criterion on which raters were the most unwilling to award high scores was Style. Vocabulary was the criterion which raters assessed most leniently. Grammar and Task achievement are between the two, and are highly similarly interpreted by raters in terms of leniency. Although assessors exhibit a slightly different level of harshness in relation to the four criteria, no criterion acts as a misfitting or overfitting one, as infit mnsq ranges between the acceptable levels of .75 and 1.18. It should be noted that the Zstd values are outside the normal range, but these can be ignored (Linacre, 2003-6; McNamara, 1996) if the mean square values are acceptable.

Apart from the assessment criteria, rater and rating scale interaction also involves the use of actual scores and how raters segment the performance continuum into 6 segments, to use scores from 0 to 5. The next table gives an example how one rater perceives the different scoring categories.

Table 10 Sample category statistics report

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5		at	Prob
0	8	1%	1%	-5.19	-6.37	2.0			(-8.15)		low	low	100%
1	70	10%	12%	-3.84	-3.75	1.0	-7.06	.45	-5.45	-7.14	-7.06	-7.09	71%
2	208	31%	43%	-1.77	-1.78	1.1	-3.86	.16	-2.42	-3.88	-3.86	-3.87	67%
3	289	43%	86%	.58	.57	1.0	-1.03	.11	1.31	-.89	-1.03	-.98	84%
4	84	13%	99%	4.43	4.49	1.1	3.72	.18	5.94	3.70	3.72	3.70	83%
5	9	1%	100%	6.83	7.20	1.1	8.23	.40	(9.30)	8.24	8.23	8.23	100%

Table 10 monitors the use of the scale structure by Rater 13. The first part of the table gives the frequencies of the individual scores used by the rater. It appears that the extreme categories are relatively rarely used, and the scores in the middle categories are more frequently targeted. In the fifth column the average measures increase with the category score which means that higher proficiency corresponds to higher scores. The category steps or thresholds are also clearly differentiated in the Steps calibration column. This means that each category has its unique position on the 0-5 continuum. It is interesting to note that this continuum is divided differently by different raters, and the category thresholds are located at slightly different points of the scale. In other words, for example category 4 might have in part the same meaning for two raters, but a minor part of their interpretation of category 4 is idiosyncratic. The distinct categories are also graphically represented in the output. It can be seen in Table 10 that the six categories are well functioning and the intersection of the different category curves are at the points displayed in Figure 9.

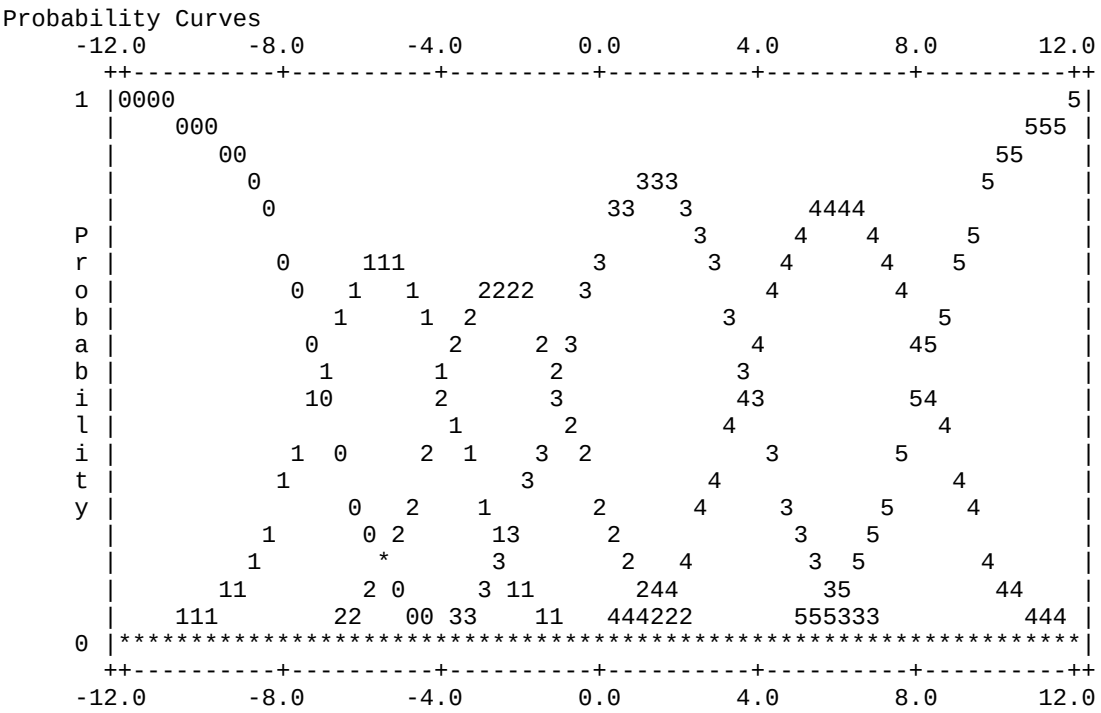


Figure 9 Probability curves for the scale steps

If a vertical line is drawn from the intersection points of the probability curves, it can be clearly seen that each category has its unique position on the x axis. Although the segments cut out of the logit axis are not completely identical in length, for example Category 3 is larger than Category 2, but the distances are still enough to constitute different categories. The step thresholds are to increase by at least 1.4 logits (Bond & Fox, 2001), which is the case in Rater 13's probability curve, but the increase should not exceed 5 logits (Linacre, 1999).

Two final tables are discussed next, both of which identify unusual interactions between raters, criteria and candidates. Each dataset as a rule includes some misfit to the model. Table 11 reports unusual rating patterns that cause misfit in the data.

Table 11 Unexpected responses

Cat	Step	Exp.	Resd	StRes	Num	Student	N	Rater	N	Criteria
5	5	3.3	1.7	3	13	040112	2	Rater13	4	Grammar
1	1	2.7	-1.7	-3	22	040121	2	Rater13	1	Task achievement
0	0	1.8	-1.8	-3	94	040192	2	Rater13	4	Grammar
4	4	3.0	1.0	3	142	040240	3	Rater7	3	Style
4	4	3.0	1.0	3	164	040262	3	Rater7	4	Grammar
2	2	3.1	-1.1	-3	170	040268	3	Rater7	4	Grammar
1	1	2.8	-1.8	-3	214	040312	4	RaterG2	1	Task achievement
4	4	1.9	2.1	3	259	040357	1	RaterG4	4	Grammar
0	0	1.9	-1.9	-3	259	040357	2	Rater12	3	Style
0	0	2.0	2.2	-3	265	040363	2	Rater13	4	Grammar
Cat	Step	Exp.	Resd	StRes	Num	Student	N	Rater	N	Criteria

The first two columns refer to the categories, that is, the actual scores which the candidate identified by a number and a code in the sixth and seventh columns of the table obtained on the criterion specified in the last column, as given by the rater identified in the Rater column. The third column specifies what the expected score would have been based on the current estimations, and the residual shows the difference between the expected and the observed score. This table is highly informative about rater consistency. A recurring rater and criterion interaction might indicate the existence of bias, especially if the direction of the misfit seems to be constant. In this dataset Rater13 exhibits the highest number of unexpected patterns. Out of the four unexpected responses three are related to the grammar criterion. There does not seem to be a consistent pattern in how Rater13 treats the grammar criterion. Twice out of the three unexpected cases she awarded lower scores than would have been expected, but for candidate 040112 she gave 5 instead of the expected 3.3. A general pattern in the unexpected responses might indicate that the rater is biased and treats a criterion differently. It should be noted that the only criterion in this dataset that did not yield unexpected responses was vocabulary which seems to suggest that this is the most comprehensible criterion, and it is interpreted by raters in a similar fashion. Finally, the

true bias terms are identified by Table 12. The differences between the observed and the expected scores are transformed into bias measures: the bigger the difference, the higher the bias size, which is expressed in log odd units.

Table 12 Bias analysis report showing rater and criteria interaction

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Model S.E.	t	Infit MnSq	Outfit MnSq	Sq	N	Rater	measr	N	Criteria	measr
295	283.2	100	.12	.55	.22	2.57	1.0	1.0	11	3	Rater7	-.73	3	Style	.54
359	337.7	134	.16	.45	.15	3.10	1.0	1.0	1	1	RaterG4	.74	1	Task achievement	.10
453	438.5	167	.09	.33	.15	2.18	1.3	1.3	14	2	Rater13	.93	4	Grammar	-.11
311	305.9	100	.05	.24	.22	1.09	.7	.6	7	3	Rater7	-.73	2	Vocabulary	-.53
410	406.4	133	.03	.12	.18	.64	.9	.9	16	4	RaterG2	-.94	4	Grammar	-.11
299	296.9	100	.02	.10	.22	.46	1.0	1.0	15	3	Rater7	-.73	4	Grammar	-.11
402	400.2	133	.01	.06	.18	.33	1.0	.9	4	4	RaterG2	-.94	1	Task achievement	.10
367	366.7	134	.00	.01	.15	.04	.8	.8	5	1	RaterG4	.74	2	Vocabulary	-.53
315	316.8	134	-.01	-.04	.14	-.26	1.0	1.0	9	1	RaterG4	.74	3	Style	.54
455	456.8	167	-.01	-.04	.15	-.28	.8	.8	6	2	Rater13	.93	2	Vocabulary	-.53
385	386.5	133	-.01	-.05	.18	-.27	.7	.7	12	4	RaterG2	-.94	3	Style	.54
425	429.5	167	-.03	-.10	.15	-.67	1.2	1.2	2	2	Rater13	.93	1	Task achievement	.10
416	419.4	133	-.03	-.11	.18	-.61	.7	.7	8	4	RaterG2	-.94	2	Vocabulary	-.53
401	409.4	167	-.05	-.18	.15	-1.23	1.0	1.0	10	2	Rater13	.93	3	Style	.54
327	347.1	134	-.15	-.43	.15	-2.94	1.3	1.2	13	1	RaterG4	.74	4	Grammar	-.11
274	292.6	100	-.19	-.88	.22	-4.03	1.0	1.0	3	3	Rater7	-.73	1	Task achievement	.10

The bias analysis report summarizes the rater and rating scale category interaction for all raters and all categories. As it is apparent from the table, even those bias terms are included which are not significant. The first two columns show the observed and the expected scores given by a certain rater on a given criterion. The third column displays the number of observations, that is, the number of papers marked. The average difference between the observed and the expected score is shown in column 4. This difference is transformed into logit value in the next column (column 5) followed by the error of the bias estimate in column 6. In the next column the bias estimates are converted into t-scores. Whereas a t-score lower than -2.0 shows that the particular rater scores the criterion in issue more harshly than other criteria, a t-score higher than +2.0 indicates that the criterion is scored more leniently than the others. The infit mean square statistics show how consistent the bias is for the rater across all candidates. In the dataset investigated as a sample, 5 significant biases can be detected. Rater7 is

lenient on the Style criterion (t = 2.57) and harsh on the Task achievement criterion (t = -4.03). RaterG4 is lenient on the Task achievement criterion and harsh on the Grammar criterion (t = -2.94). Finally, Rater13 is lenient on the Grammar criterion (t = 2.18).These data show that no general pattern can be detected in raters' handling the different criteria, as they all have their personal interpretation of the criteria and act accordingly. In addition, it should be noted that these observations regarding the biases are related to these data and cannot be generalized and claimed to be a permanent characteristics of the rater. Figure 10 gives a graphic representation of raters' personal understanding and use of the rating scale.

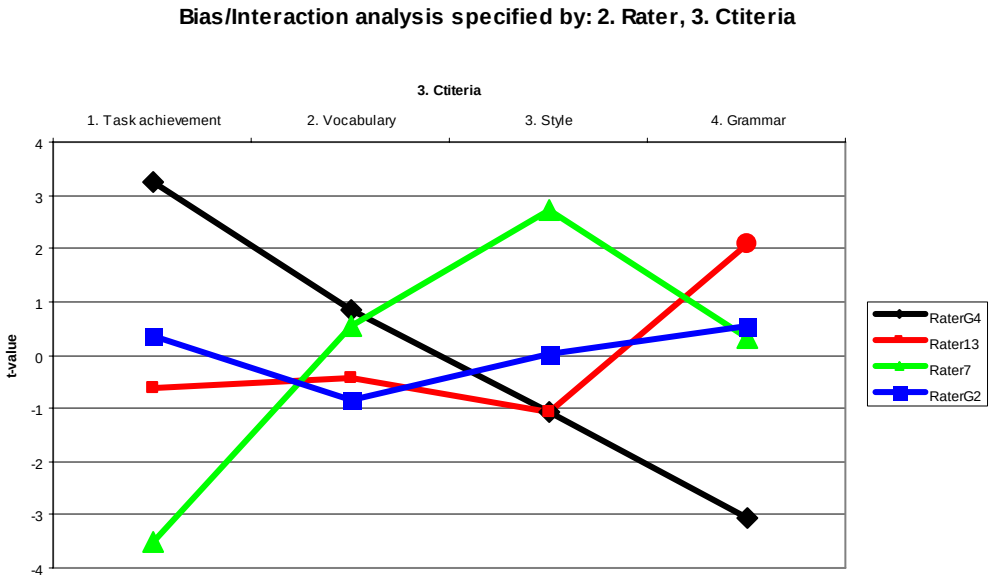


Figure 10 Rater-rating scale interaction

The types of statistical data in Tables 7-12 shown and interpreted as examples will be investigated in Study 1 in order to obtain information about the functioning of the rating scale, the criteria and the categories as well as rater behaviour.

5.2 Rater misbehaviour

5.2.1 Rater idiosyncrasies

In discussing rater idiosyncrasies, three aspects of rater behaviour were investigated. Firstly, one of the most commonly researched areas in subjective rating, rater leniency and harshness are explored and discussed. This characteristics is expressed numerically in logits, the common measurement unit of IRT. Secondly, rater inconsistencies are studied with the help of analysing infit statistics. Rater inconsistencies can also be expressed numerically. As it has been proposed in the previous section, infit mean square values express the compatibility of the data with the model (Bond & Fox, 2001), with 1 as the expected value. As Wright and Linacre claim, "though the ideal for measurement construction is that data fit the Rasch model, all empirical data departs from it to some extent " (1994, p.370). Fit values over 1 indicate more variation in the observed values than expected by the model, and values lower than 1 show less variance than expected by the model.

Finally, inter-rater agreement is studied based on expected and observed agreement according to model predictions. When two raters are applied in an assessment situation, neither complete agreement nor complete disagreement is ideal. According to Linacre (2003-6), if the observed agreement largely exceeds the expected agreement, this indicates "forced agreement", an amount of redundancy that the model would not expect. In other words, raters behave like scoring machines which, in a way, makes double scoring unnecessary. If the observed agreement, however, is much lower than the expected agreement, there is a substantial difference between the two raters' value judgements. These three types of indices are collected from the Raters Measure Report table in the FACETS output and tabulated in Table 13. In the study I investigated the differences in leniency teachers of German and teachers of English

show when marking writing tasks. Three examination periods were studied and in each period both German and English markers' leniency was compared. The raters' leniency was compared both for the English and the German group in the periods under investigation in years 2004, 2005 and 2006. As it might have been expected, there are some differences between raters' leniency but these differences are not substantial. The size of the difference between the most and the least lenient rater in most cases does not exceed one logit.

Table 13 Summary of rater characteristics expressed in figures across six datasets

Design	Rater	Measure	Infit msq	Observed	Expected
2004_German_1					
	RaterG2	-.94	.85	68.4	56.1
	Rater7	-.73	1.01	72.8	58.0
	RaterG4	.74	1.03	50.7	48.0
	Rater13	.93	1.08	51.6	48.0
2004_German_2	RaterG1	-.56	.86	50.4	50.5
	RaterG3	-.19	1.24	46.1	52.4
	Rater4	-.76	.94	57.0	47.6
2004_English	RaterE3	-.35	1.2	26.1	32.7
	Rater12	-.26	.73	36.4	33.8
	RaterE2	-.12	.97	33.8	34.0
	Rater5	-.03	1.11	24.0	32.0
	Rater2	-.14	1.03	31.8	33.4
	Rater3	.62	.66	24.0	31.0
2005_English	RaterE2	-.76	1.11	67.1	51.2
	Rater12	-.57	1.10	67.5	45.5
	RaterE4	-.38	.84	70.0	51.3
	Rater8	1.7	1.27	30.9	32.7
2005_German	RaterG4	-.50	1.53	38.5	44.2
	RaterG2	-.15	.87	60.2	51.0
	Rater13	.65	.84	51.7	48.3
2006_English_1	Rater2	-.12	1.09	51.1	45.7
	RaterE1	.12	.90	51.1	45.7
2006_English_2	RaterE8	-.78	1.03	57.6	62.1
	RaterE9	-.48	.74	74.9	61.0
	RaterE7	-.32	1.08	55.0	60.4
	RaterE6	.61	1.06	75.0	61.1
	RaterE5	.96	1.00	53.6	59.3

The first column shows the sources of the data, the second includes the code of the rater. Next, in the Measure column, rater strictness is shown in logit values. The lower the logit value, the more lenient the rater, and the higher the logit value, the harsher the rater. The final two columns present the observed scores, that is, the actual scores given by the rater and scores expected by the model. Rater leniency is indicated by observed scores higher than the expected. Conversely, a lower observed score than expected by the model means that the rater is harsher than what the model expects. Even if differences in overall leniency seem insignificant, they provide useful information for the arrangement of the rating pairs. The infit values in Table 13 with one exception do not attest to significant misfit (infit values over 1.5), which means that the raters under investigation show a reasonable consistency in their rating behaviour. Neither is there overfit in the data (fit values below .5) which means that there is sufficient variability in the scores raters used to assess writing performances.

The rater agreement patterns in Table 14 show an interesting picture. There are no noticeable differences between the expected and observed agreement cases. On the whole, it can be concluded that the observed exact agreement cases (column 3) exceed the expected agreement cases more often than remain below them. The differences between the percentages indicating expected and observed cases are small, especially when the number of expected agreement cases is lower than that of the observed agreement cases. The Rasch Kappa values are all smaller than +/- 0,00. This general pattern suggests a high degree of agreement between raters especially considering the amount of data in which agreement was sought to be achieved.

Table 14 Summary of rater agreement indices

Design	Inter-rater agreement opportunities	Exact agreement	Expected agreement
2004_German_1	1068	59.6 %	52.9 %
2004_German_2	508	50.4 %	50.5 %
2004_English	1904	31.0 %	33.3 %
2005_English	1855	67.2 %	50.0 %
2005_German	660	51.7 %	48.3 %
2006_English_1	800	51.1 %	45.7 %
2006_English_2	1728	64.3 %	60.7 %

5.3 Rating scale use

5.3.1 Rating scale criteria

Rating scales cannot be studied independently of raters, therefore an investigation of the functioning of the steps and the criteria will follow. First, the four categories of the rating scale: Task achievement, Style, Vocabulary and Grammar, are investigated in detail. The results of the FACETS analyses for each examination period under investigation will be tabulated and compared to identify any possible recurring patterns. If a tendency can be observed concerning the leniency and harshness of raters' category interpretation, it might indicate an unwanted weighting of a certain criterion by the raters. Table 15 presents raters' assessment category use in the six different rating designs.

Table 15 Summary of rating scale category functioning

Design	Criterion	Measure	Infit	Estimated discrimination
2004_German_1				
	Vocabulary	-.51	.78	1.22
	Grammar	-.10	1.16	.86
	Task achievement	.09	1.11	.88
	Style	.52	.94	1.05
2004_German_2				
	Style	-.81	1.03	.98
	Vocabulary	.01	.85	1.14
	Task achievement	.04	1.13	.89
	Grammar	.76	.97	1.02
2004_English				
	Grammar	-.10	.93	1.06
	Task achievement	-.08	1.11	.88
	Style	.06	1.03	.98
	Vocabulary	.12	.92	1.07
2005_English				
	Style	-.40	.89	1.09
	Grammar	-.34	.68	1.32
	Vocabulary	.00	.60	1.35
	Task achievement	.74	1.76	.24
2005_German				
	Vocabulary	-.28	.74	1.25
	Grammar	1.18	1.12	.89
	Task achievement	.14	1.36	.75
	Style	.32	.78	1.18
2006_English_1				
	Vocabulary	-.22	.70	1.30
	Task achievement	-.21	1.15	1.02
	Grammar	-.21	.86	1.14
	Style	.64	1.26	.62
2006_English_2				
	Task achievement	-1.43	1.14	.83
	Style	.01	.82	1.12
	Vocabulary	.24	.78	1.16
	Grammar	1.18	1.10	.91

The first column in the table displays the different examination periods in which the scores of writing tasks were collected. The second and the third columns specify the criteria and show the leniency or harshness raters displayed towards the rating criterion. In other words, it is shown which criterion was treated in a more generous way by raters, which was the one where it was easy for candidates to get a higher score on and which was the one in the use of which raters tended to be stricter, and it was more difficult for candidates to get a higher score on. The third column shows the logit values, that is, the "difficulty" value of the criteria. The lowest the logit value of the criterion, the easier to obtain a higher score on it.

First of all, it should be noted that the range of logit values of the criteria is rather restricted, as they are not spread out widely. The logit values are all clustered around zero, with no measure exceeding 1.5 or going below -1.5. In the three years 2004, 2005 and 2006, and for the two language groups, English and German, under investigation no systematic deviation can be detected in the use of the criteria. A rank ordering of the criteria, with 1 as the most leniently scored criterion and 4 as the most strictly scored criterion, shows that the differences average out across the examination periods (Vocabulary $\overline{X}$ =2,14, Task achievement and Grammar $\overline{X}$ = 2,57 and Style $\overline{X}$ =2,7). One has to note, however, that averaging is a fair procedure only in trying to capture consistent bias in investigating results for longer periods. For one special examination period substantial differences associated with certain criteria do indeed matter. According to the results, however, neither the English nor the German group shows a systematic differential treatment of any of the criteria.

All these data suggest that the criteria are not functioning differentially, and that raters do not apply criteria differentially. Although there are minor differences in the

leniency or harshness with which they treat a certain criterion, these differences are neither systematic nor large enough to lead to systematic measurement error.

The infit data reveal how consistently raters apply the criteria. With the acceptable range defined earlier, it seems that only in one case does the infit value exceed the acceptable figure of 1.5: in the 2005_English design the infit value for task achievement is 1.76. The expectation of infit, the information-weighted mean-square fit statistic is 1, and values over 1 indicate unmodelled excess variation or unexpected irregularities. At this point it seems necessary to reiterate how infit should be interpreted. Linacre (2003-6, p.149) claims that although 1 is the expected value for infit mean square, minor deviations from the expected value do not degrade the measurement process. He considers mean square values between 1.5 and 2.0 "unproductive for measurement but not degrading" (ibid, p.149). Thus, the infit value of 1.76 does not seem to give reason for major concern. The final column shows how well the items discriminate: in this case the number shows the discriminating power of the criteria. The expected value is 1, negative figures indicate reverse discrimination.

The data in the table show a discrimination value around 1, with one exception. It seems that Task achievement in the 2005_English design has a very low discrimination value which might be related to the problem suggested by the relatively high infit value. This result might indicate further problems, such as lack of variance of the awarded scores, for example due to the overuse of one category. Further results explaining rater behaviour and unusual ratting patterns will clarify the reason for the high infit value and the low discrimination. Apart from this case, namely the differential functioning of the Task achievement criterion in the 2005_English design, it seems appropriate to say that no systematic deviation in the application of the rating scale categories can be detected. They are all used independently, they discriminate

appropriately and no systematic overuse or underuse of any one category has been empirically confirmed.

5.3.2 Rating scale categories

Next, the six steps, 0 through 5, of the rating scale were examined. The purpose of this investigation was to confirm the appropriate use of the six categories and to provide empirical evidence to prove that it is possible to differentiate between different levels of achievement with the application of a six-point scale. It was expected that lower categories were associated with lower measures and higher categories with higher measures. In our case a lower category, for example 1 is associated with less ability than 3. If this is not the case and a specific category is associated with lower measure than the previous lower category, the figures are flagged with an asterisk in the output table. The next table summarizes the most relevant data from the FACETS output file and provides information about the use of the different categories of scores.

Table 16 Summary of scale step statistics

		Category counts used	%	Average measure	Expected measure	Outfit MnSq	Step calibration
2004_German_1							
	0	28	1	-4.93	-5.13	1.2	
	1	185	9	-3.66	-3.58	.9	-6.28
	2	587	27	-1.49	-1.51	1.0	-3.73
	3	887	42	1.02	1.02	1.0	-.73
	4	371	17	4.22	4.26	1.0	3.47
	5	78	4	7.02	6.91	.9	7.27
2004_German_2							
	0	9	1	-5.54	-5.95	1.7	
	1	36	4	-3.29	-3.27	1.0	-6.14
	2	266	26	-.70	-.64	.9	-3.84
	3	428	42	1.41	1.38	1.0	-.13
	4	207	20	4.18	4.17	1.0	3.45
	5	70	7	6.81	6.82	1.0	6 .66
2004_English							
	0	44	2	-1.09	-1.42	1.4	
	1	224	8	-.67	-.59	1.0	-2.60
	2	580	20	.10	.08	1.0	-1.20
	3	947	33	.72	0.75	.9	-.80
	4	711	25	1.44	1.40	1.0	1.36
	5	334	12	2.16	2.17	1.0	2.52
2005_English							
	0	131	4	-4.38	-5.27	2.2	
	1	674	18	-3.93	-3.69	.8	-6.17
	2	782	21	-1.57	-1.53	.7	-2.77
	3	1102	30	-.92	.85	.8	-.72
	4	819	22	4.04	4.00	1.0	2.65
	5	179	5	6.65	6.75	1.1	7.01
2005_German							
	0	38	3	-4.00	-4.70	1.7	
	1	172	13	-3.88	-3.59	.7	-5.68
	2	346	26	-1.97	-2.05	.9	-3.57
	3	505	38	.49	.52	1.0	-1.25
	4	219	17	4.05	4.02	.9	3.04
	5	40	3	7.24	7.13	.8	7.46
2006_English_1							

	0	95	6	-3.83	-4.19	1.7	
	1	136	9	-3.37	-3.00	.9	-4.00
	2	250	16	-.93	-1.02	1.1	-2.68
	3	378	24	1.18	1.25	.8	-.30
	4	479	30	3.47	3.45	.9	2.17
	5	246	16	4.78	4.74	1.0	4.18
2006_English_2							
	0	25	1	-10.32	-10.69	1.2	
	1	271	8	-6.63	-6.67	1.1	-11.15
	2	787	23	-1.19	-1.16	1.0	-4.88
	3	1257	37	2.78	2.80	.9	.38
	4	855	25	7.23	7.19	1.0	5.33
	5	237	7	10.60	10.60	1.0	10.32

After the design is specified in the first column, the second column presents the six categories of the scores from 0 to 5. The next two columns in the table present the frequency of the use of the score category and the percentage they represent in the score category use. The numbers in the first four columns in the table indicate the following: in the 2004_German design zero scores were awarded on 28 occasions which represents a total of 1% of all score category use. The next two columns give the actual and the expected measure associated with the given score category. It is expected that a lower score category corresponds to lower measure as it stands for a lower level of proficiency. The average measure for the category represents the average ability estimate for the candidate getting the score. In addition, it is expected that there should be a monotonically increasing level of measures as the categories are growing. If this is not the case, the numbers are flagged. It is clearly visible that there are no flagged data in the table, which means that the score categories are used in a consistent way, and higher proficiency is rewarded with a higher score. This might appear an evident presupposition, yet this is not necessarily always the case. The purpose of a validation study should be the investigation of how the steps of a scale are functioning in the case

of an operational assessment scale. The category counts used and the percentages indicate that with one exception 3 is the most frequently occurring score with the highest category counts and percentages. This resembles a quasi normal distribution with extremes scores appearing with a much a lower frequency. The actual average measures and the expected measures display an evenly increasing level of measures confirming the appropriate functioning of the scale, namely that higher proficiency (measure) is rewarded with a higher score (category).

In this FACETS output, no infit mean square values are shown, only outfit mean square is reported. The expected value is 1, with larger numbers for the extreme categories. The outfit mean square values show a good fit, they are all around the expected value of 1, except for some of the extreme categories. There seems to be one unusually high outfit mean, 2.2 for a zero score in the 2005_English design. In the previous analysis the same design indicated some problem with the Task achievement criterion, which showed little discrimination. From the results that have emerged so far it seems that the zero category with the Task achievement criterion shows some irregular pattern in the 2005_English design. The final column presents data about the step calibrations or the thresholds in the scale. Values in this column, like the average measures, are also expected to show a steadily growing tendency. As it has already been pointed out, there should be a difference of at least 1.4 logits (Bond & Fox, 2001), but the differences should not exceed 5 logits (Linacre, 1999). The data in the table show that the rating scale categories are part of a well-functioning measurement tool. There seem to be two cases where the values are outside the set boundaries. In the 2004_English design categories 2 and 3 are slightly closer to each other than expected (-1.2 and -.08). This does not seem to distort the scale as the increase in the measures is still evident, and the outfit value does not indicate any problems either. The information

on the functioning of the scale should be used in combination with other data to get a complete picture (Bond & Fox, 2001), thus a minor distortion from the norm in one parameter only does not necessarily indicate a serious flaw of the scale. The other case, where there appears to be a difference in the step calibrations bigger than expected, is in the 2006_English_2 design. The differences exceed 5 logits at the ends of the scale. It is not uncommon to encounter extreme values at the end of the scales as the IRT application produces the most stable estimates around the middle of the scale.

On the whole, it can be concluded that the rating scale measures and steps are functioning well, the six steps are clearly identifiable, well differentiated, and they represent clearly different levels of proficiency and achievement in the exam. The category frequencies attest to the need of six different category scores. The application of the six steps to differentiate between proficiency levels appears to be empirically confirmed by the analysis.

5.3.3 Unexpected responses

FACETS can identify unexpected responses which show a marked irregularity within the dataset. By marked irregularity unexpectedly large residuals are meant. Even when the data fit the model, there might be some items with larger residuals than the model expectations. The unexpected responses output table makes a list of the unexpected responses, defines the expected values and gives the actual observation. These unexpected responses might be due to clerical error but also to “idiosyncratic, off-variable” behaviour. In rating scale use, these unexpected responses may indicate unusual rater behaviour which, if showing a recurrent pattern, can indicate systematic rater error. The unexpected responses are presented in two tables. First, in Table 16 the unexpected responses that were indicated by the programme are summarized. Second, a

separate table is produced for the 2005_English design which does indeed indicate some form of systematic deviation.

Table 17 Summary of unexpected responses

Design	Category	Expected	Residual	Rater	Criteria
2004_German_1					
	5	3.3	1.7	Rater13	Grammar
	1	2.7	-1.7	Rater13	Task achievement
	0	1.8	-1.8	Rater13	Grammar
	4	3.0	1.0	Rater7	Style
	4	3.0	1.0	Rater7	Grammar
	2	3.1	-1.1	Rater7	Grammar
	1	2.8	-1.8	RaterG2	Task achievement
	4	1.9	2.1	RaterG4	Grammar
	0	1.9	-1.9	Rater13	Style
	0	2.2	-2.2	Rater13	Grammar
2004_German_2	0	2.2	-2.2	RaterG1	Task achievement
	4	2.2	1.8	RaterG1	Vocabulary
	2	3.8	-1.8	RaterG3	Task achievement
2004_English	0	4.0	-4.0	Rater1	Task achievement
	0	3.5	-3.5	RaterE2	Task achievement
	0	2.7	-2.7	RaterE2	Task achievement
	0	3.0	-3.0	Rater1	Task achievement
	1	3.7	-2.7	RaterE3	Style
	1	3.7	-2.7	Rater1	Grammar
	1	4.0	-3.0	Rater1	Style
2005_German	0	2.2	-2.2	RaterG2	Task achievement
	4	1.9	2.1	RaterG2	Grammar
	0	2.2	-2.2	RaterG4	Task achievement

	0	2.2	-2.2	RaterG4	Task achievement
	0	1.9	-1.9	RaterG4	Task achievement
2006_English_1	0	2.9	-2.9	RaterE1	Task achievement
	2	4.3	-2.3	Rater2	Style
	2	4.5	-2.5	Rater2	Style
	0	2.2	-2.2	RaterE1	Task achievement
	0	2.3	-2.3	Rater2	Task achievement
	5	3.0	2.0	Rater2	Vocabulary
	5	2.7	2.3	Rater2	Style
	2	4.0	-2.0	RaterE1	Style
	0	2.3	-2.3	RaterE1	Task achievement
	2	3.9	-1.9	Rater2	Style
	2	4.2	-2.2	Rater2	Style
	4	1.8	2.2	Rater2	Grammar
	2	4.2	-2.2	Rater2	Style
	3	.8	2.2	Rater2	Grammar
	2	3.9	-1.9	RaterE1	Task achievement
2006_English_2	4	1.9	2.1	RaterE6	Grammar
	1	3.2	-2.2	RaterE6	Style
	4	5.0	-1.0	RaterE6	Task achievement
	5	3.2	1.8	RaterE7	Style
	0	1.0	-1.0	RaterE9	Task achievement
	0	1.0	-1.1	RaterE6	Task achievement
	2	.9	1.1	RaterE6	Vocabulary
	2	1.0	1.0	RaterE9	Vocabulary
	4	4.9	-.9	RaterE9	Task achievement
	2	1.0	1.0	RaterE9	Grammar
	2	1.0	1.0	RaterE9	Grammar
	0	1.1	-1.1	RaterE6	Grammar
	4	2.4	1.6	RaterE6	Vocabulary

2	3.6	-1.6	RaterE6	Task achievement
0	1.0	-1.0	RaterE5	Task achievement
2	3.7	-1.7	RaterE7	Task achievement
0	1.0	-1.0	RaterE7	Grammar

Table 17 presents the unexpected responses in the six designs investigated, and one design will also be examined separately. The Category column presents the score that was actually awarded, and it is followed by what score had been expected by the model. The Residual column displays the difference between the expected and the observed score: the expected score is subtracted from the observed score. If the figure is negative, the candidate was marked down, in other words, the candidate was disadvantaged, as his actual score was lower than would have been expected in all probabilities. A positive residual indicates that the candidate received a higher score than would have been expected by the model. The final two columns identify the raters and the criteria, respectively. This table might identify biases in the traditional sense of the word if a recurring pattern can be detected in the dataset and may also reveal unsystematic rater behaviour.

The frequency of the scale categories, in other words, the scores in the second column clearly indicate that the most problematic score is zero. Almost half of all the unexpected responses, namely 21 cases, are associated with the zero score category. Score category 2 seems to be the second most problematic, as it yields unexpected patterns in 15 cases. The other categories seem to result in unexpected responses much less frequently: score 1 on six occasions, score 3 only once, score 4 on ten occasions and score 5 on four occasions. As the zero score category is one end of the scale, it seems evident that candidates are often assessed below their merits. The residuals presented below in Table 18 will also reveal that they receive zero scores instead of a

well deserved 1 or 2, or an even higher score. The frequency of criteria in the unexpected responses shows that the most problematic criterion is Task achievement, which is the source of unexpected responses in 24 cases. Grammar and style come next with 15 and 13 unexpected occurrences. Out of the four assessment criteria, vocabulary seems to be the least common source of unexpected scores. This criterion is associated with unexpected responses only on five occasions. So far the results obtained from the unexpected responses seem to suggest that the zero score category and the Task achievement criterion might be the source of rater biases.

In order to obtain a more accessible picture of which criteria and which score categories result in unexpected responses, Table 18 arranges the residuals, the difference between the expected and the observed score, with the corresponding criteria in an ascending order. At the top of the table are the largest negative residuals with the corresponding assessment criteria. These are the ones in the use of which candidates received lower scores than they should have received and thus were disadvantaged. At the other end of the list are the criteria with the largest positive residuals. In other words, these are scores in the case of which candidates were unduly advantaged. This table also casts light on the size of the deviations.

Table 18 Residuals in an ascending order with the associated score categories and criteria

Category	Residual	Criteria
0	-4	Task achievement
0	-3,5	Task achievement
0	-3	Task achievement
1	-3	Style
0	-2,9	Task achievement
0	-2,7	Task achievement
1	-2,7	Style
1	-2,7	Grammar
2	-2,5	Style

2	-2,3	Style
0	-2,3	Task achievement
0	-2,3	Task achievement
0	-2,2	Grammar
0	-2,2	Task achievement
0	-2,2	Task achievement
0	-2,2	Task achievement
0	-2,2	Task achievement
0	-2,2	Task achievement
2	-2,2	Style
2	-2,2	Style
1	-2,2	Style
2	-2	Style
0	-1,9	Style
0	-1,9	Task achievement
2	-1,9	Style
2	-1,9	Task achievement
0	-1,8	Grammar
1	-1,8	Task achievement
2	-1,8	Task achievement
1	-1,7	Task achievement
2	-1,7	Task achievement
2	-1,6	Task achievement
2	-1,1	Grammar
0	-1,1	Task achievement
0	-1,1	Grammar
4	-1	Task achievement
0	-1	Task achievement
0	-1	Task achievement
0	-1	Grammar
4	-0,9	Task achievement
4	1	Style
4	1	Grammar
2	1	Vocabulary
2	1	Grammar
2	1	Grammar
2	1,1	Vocabulary
4	1,6	Vocabulary
5	1,7	Grammar
4	1,8	Vocabulary
5	1,8	Style
5	2	Vocabulary
4	2,1	Grammar
4	2,1	Grammar
4	2,1	Grammar
4	2,2	Grammar
3	2,2	Grammar
5	2,3	Style

The frequencies of the individual categories have already been discussed. At this point, the combination of the score category and the criteria together with the size of the residual, which in simple terms can be regarded as the deviation, is to be interpreted. One important issue should be established right at the outset. Although Table 18 seems to present a long list of unexpected responses to rater misbehaviour and differential rater and rating scale functioning, with such a large dataset this amount of unexpected observations cannot be considered highly significant. The 2011 papers, each rated by two raters on four criteria result in 16088 scores out of which 57 appeared as unexpected. Expressed in percentages, the unexpected scores amount to 0.003 % of all cases. It should also be pointed out that the dataset comes from double scoring before agreement is reached between the two raters concerning the final score of the candidate. Nevertheless, these discrepancies cannot be ignored, and as the aim of the project is the validation of a rating scale and the rating process, the exploration of these minor discrepancies and a conscious effort at eliminating them might largely contribute to the appropriate functioning of the rating scale and the rating process.

It seems from the first part of the table with the negative residuals, where candidates were scored down, that the dominant score category is zero and the dominant criterion is Task achievement in the analysis of the unexpected residuals. The three variables, the score category, the residual and the criterion category are ordered according to the size of the residual, which clearly indicates an emerging pattern: there seems to be a tendency to award zero scores on the Task achievement criterion. It also appears from the data that the differences between expected and observed scores exceed one score in the majority of the cases. In practical terms this would mean 2, 3 or even 4 points loss for a candidate on one criterion only. The positive residuals mean that the candidate was awarded a higher score than would have been expected. Here the

residuals are much smaller between the observed and expected values, which means that raters' deviations which would benefit the candidate are smaller than deviations that would disadvantage them. The dominant malfunctioning category is Task achievement. Grammar and Vocabulary are the other two criteria, which result in unexpected responses that would advantage the candidates. The Style criterion does not show a consistent pattern in terms of causing advantage or disadvantage to the candidate. Based on the data in Table 18, it can be claimed that the most frequently occurring unexpected rating patterns are zero scores on the Task achievement criterion. On the other hand, slightly higher scores than the expected are given on the Grammar criterion and to a lesser extent on Vocabulary. Marking performances down, even if unintentionally, is obviously an act of disadvantaging a candidate. The data used for analysis here are not the final agreed scores that a candidate eventually received for his/her writing performance. Although these results are highly informative, nevertheless neither the amount of unexpected response patterns nor the size of deviations create an impression of pursuing unfair testing practice.

5.3.4 Further exploration of the deviations

The results obtained from the analysis of the unexpected responses have called for a differential treatment of one dataset, that of the 2005_English design. The pattern that has emerged from the analyses so far is presented in the table in a detailed form. Unlike in the previous tables, here the raters are also specified. The Task achievement criterion seems to be the one treated by RaterE4 differentially, or in other words RaterE4 shows bias towards the Task achievement criterion. RaterE2 shows a very similar rater profile, s/he also produces unexpected responses on the Task achievement criterion, although in a less consistent manner. His/her value judgement fluctuates on

this criterion, and there is no consistency in the set of unexpected responses. In the majority of the cases s/he is harsh on that criterion, but on one occasion s/he appears to display a generous rater attitude on this criterion.

Table 19 Unexpected responses/ratings in the 2005 English design

Cat	Expected	Residual	Raters	Criteria
0	3.8	-3.8	RaterE2	Task achievement
4	1.1	2.9	Rater12	Task achievement
0	3.0	-3.0	RaterE2	Task achievement
0	3.0	-3.0	RaterE2	Task achievement
0	3.2	-3.2	RaterE4	Task achievement
0	3.1	-3.1	RaterE4	Task achievement
0	2.7	-2.7	RaterE2	Task achievement
1	3.2	-2.2	RaterE2	Task achievement
0	2.6	-2.6	RaterE4	Task achievement
2	3.9	-1.9	RaterE2	Task achievement
2	3.8	-1.8	Rater12	Task achievement
4	1.9	2.1	Rater8	Task achievement
1	2.8	-1.8	Rater8	Style
1	3.1	-2.1	RaterE2	Task achievement
4	4.9	-.9	RaterE4	Style
0	2.4	-2.4	RaterE2	Task achievement
0	2.3	-2.3	RaterE2	Task achievement
0	2.0	-2.0	RaterE2	Task achievement
0	2.0	-2.0	RaterE2	Task achievement
0	2.2	-2.2	RaterE2	Task achievement
0	2.4	-2.4	RaterE2	Task achievement
0	2.3	-2.3	RaterE2	Task achievement
0	2.4	-2.4	RaterE2	Task achievement
0	2.3	-2.3	RaterE4	Task achievement
5	3.1	1.9	RaterE2	Task achievement
0	2.1	-2.1	RaterE4	Task achievement
0	2.1	-2.1	RaterE4	Task achievement
0	2.1	-2.1	RaterE4	Task achievement
4	1.7	2.3	RaterE2	Style
1	2.8	-1.8	RaterE2	Task achievement
0	2.3	-2.3	RaterE4	Task achievement
1	2.8	-1.8	RaterE2	Task achievement
2	3.7	-1.7	RaterE4	Style
0	2.0	-2.0	RaterE2	Task achievement
1	3.1	-2.1	RaterE4	Task achievement
2	3.8	-1.8	RaterE4	Task achievement
1	2.9	-1.9	RaterE4	Style
1	2.8	-1.8	RaterE4	Style

5.3.5 Bias analysis

As it has been highlighted before, it is necessary to investigate all available data to be able to set up a rating scale diagnosis and make an informed decision on how the rating scale and the raters are functioning. First the raters were investigated in terms of leniency and harshness, consistency and inconsistency. Next, the elements of the rating scale were analysed with the aim of identifying possible invalidity and malfunctioning of any of the elements. As the results obtained so far suggest, the six categories of the scale are well functioning and so are the steps of the scale. The assessment criteria show little and insignificant misfit that does not pose a threat to the validity of the measurement. It is not sufficient, however, to examine the elements of the rating process in isolation, the interaction of the elements should also be a target of investigation.

Bias is defined as an interaction of elements involved in the measurement process. If specified in the model statement, special output file is produced with bias terms reported in the same frame of reference as the other elements in the analysis. Previous results have suggested that the 2005_English design warrants bias analysis.

Table 20 Bias analysis of the 2005 English rating data

Obsvd score	Exp. score	Obsvd Count	Obs-Exp Average	Bias size	Mod. S.E.	t	Infit MnSq	Raters	Criteria
57	73.6	34	-.49	-1.45	.30	-4.89	1.1	Rater8	Style
266	297.2	100	-.31	-.99	.18	-5.61	.5	Rater12	Style
276	295.5	100	-.19	-.62	.18	-3.50	.8	Rater12	Grammar
837	883.2	361	-.13	-.40	.09	-4.33	**2.0**	RaterE2	**Task achievement**
67	69.1	34	-.06	-.18	.30	-.61	.9	Rater8	Vocabulary
989	1003.9	361	-.04	-.13	.09	-1.42	.7	RaterE2	Grammar
1148	1162.3	427	-.03	-.11	.09	-1.25	.8	RaterE4	Style
998	1011.0	427	-.03	-.10	.09	-1.12	1.4	RaterE4	Task achievement
964	965.0	360	.00	-.01	.09	-.10	.7	RaterE2	Vocabulary
1110	1110.2	427	.00	.00	.09	-.02	.6	RaterE4	Vocabulary
288	285.0	100	.03	.10	.18	.54	.5	Rater12	Vocabulary
1182	1154.9	427	.06	.21	.09	2.36	.6	RaterE4	Grammar
1072	1010.1	361	.17	.56	.10	5.88	.9	RaterE2	Style
80	72.9	34	.21	.62	.30	2.10	.5	Rater8	Grammar
72	60.7	34	.33	.99	.30	3.36	1.3	Rater8	Task achievement
309	261.6	100	.47	1.52	.18	8.33	1.4	Rater12	Task achievement

Table 20 presents the result of the bias analysis. Each interaction between the raters and the assessment criteria is shown in the table, even those where no significant bias can be detected. The first two columns provide the observed and the expected scores given by the raters specified in column 9. The last column specifies the criterion which was investigated. The fourth column averages out the differences between the observed and expected scores by dividing the difference between the two by the observed count, that is, the actual number of scores given. In the fifth column this difference is shown in bias logit value followed by the error in the bias estimate. In the seventh column the bias estimates are transformed into t-values. The t-statistic is the report of a test of the statistical significance of the size of the bias. Negative bias and t-value indicate that the expected score was higher than the observed score and that the rater was stricter on the criterion than on the other criteria. A positive bias measure and t-value indicate the reverse, the observed score was higher than the expected score, and the rater was more lenient on the criterion than on the other criteria. Thus, from the data it seems that for example Rater 8 was rather harsh on the style and vocabulary criteria,

but fairly lenient on Grammar and Task achievement. Similarly, Rater12 was lenient on Task achievement and vocabulary, but harsh on Grammar and Style. Figure 11 gives a graphic representation of rater and rating scale interaction.

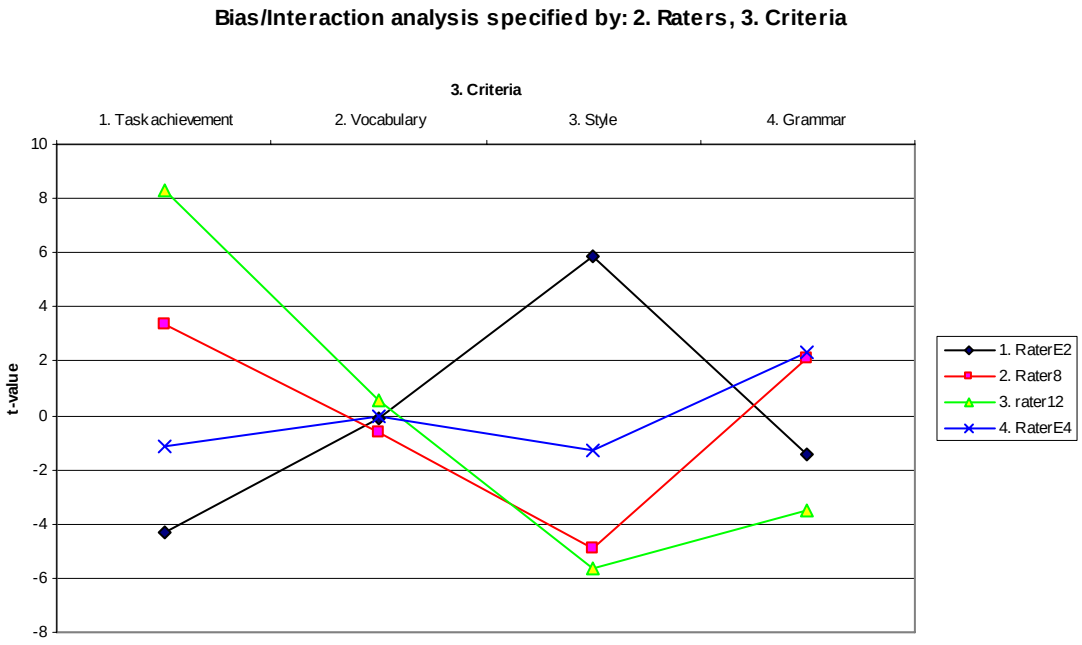


Figure 11 Graphic representation of the rater-rating scale interaction

It is apparent at first sight that raters do not attribute equal difficulty to each assessment criterion. It is interesting to investigate the graph from two different perspectives. Firstly, in terms of criterion difficulty there seems to be one evident case, that of vocabulary. Raters are all clustered around the zero point on the Vocabulary criterion, which means that they share a highly similar understanding and interpretation of this criterion and they exhibit neither lenient, nor harsh behaviour on this category. The other criteria show a less balanced pattern. Grammar shows the second most balanced spread in terms of t-values, although much less unanimous than in the case of the Vocabulary criterion. From the point of view of the raters, the graph provides

additional pieces of interesting information. Rater8 and Rater12 exhibit major differences in their treatment of the assessment criteria, whereas RaterE2 and RaterE4 seem less biased towards any of the categories. The graphs for Rater8 and Rater12 display differences between their treatment of the criteria, and it is interesting to observe that they share a similar understanding of the criteria as the two lines run in a parallel fashion. For these two raters, Task achievement is the most leniently scored criterion. On the other hand, they are equally harsh on the Style criterion, which is represented by the lowest values and lowest points in the graph. This similarity between the two raters' treatment of the criteria does not, however, apply to the Grammar criterion. Whereas Rater12 is rather strict on Grammar, Rater8 is the most lenient on this criterion among all four raters.

5.4 Conclusion

Whereas the data analysed provide reassuring feedback on the operation of the rating scale and the raters, minor discrepancies could still be detected. It seems that raters all have their personal understanding and interpretation of the rating scale, but these differences average out in the long run. It is reassuring that no regular pattern or consistent deviation could be detected in the datasets. It is, however, important to monitor rater behaviour as individual deviations and seemingly insignificant distortions should also be paid due attention.

The results seem to confirm earlier findings, according to which raters have their idiosyncratic rating patterns (Wolfe, 2004). Whereas raters are different in terms of leniency and severity (e.g. Kondo-Brown, 2002; Lee & Kantor, 2003), they appear to be consistent in their overall severity (Eckes, 2005; O'Neill & Lunz, 1995). The bias analysis investigating a hypothesized unbalanced use of assessment criteria also yielded

results similar to earlier studies (Kondo-Brown, 2002; O'Sullivan & Rignall, 2001): although it is possible to identify bias patterns in markers' rating data, no systematic or regular biases could be detected. This finding, however, contradicts Lee and Kantor's (2003) and McNamara's (1996) suggestion according to which raters appear to focus on grammar, and the language use criterion is more dominant than the other criteria.

The implications of the analysis concern chiefly issues related to the organization of test marking sessions. Firstly, the selection and pairing of raters should be informed by rater profiles based on empirical data rather than intuitions. Secondly, rater biases, as Elder (2005), and Lumley and McNamara (1995) also found, can be reduced by training. Even if raters are not interchangeable, neither can training altogether eliminate the differences between them, a constant monitoring of the rating process is indispensable as "high quality ratings are essential for valid and reliable inferences about writing competence" (Engelhard, 1994, p.95).

Chapter 6: Rater Behaviour from a Qualitative Perspective: Raters' Verbal Reports

Introduction

This chapter discusses the results of Study 2, which, through mainly qualitative inquiries, sought to identify the sources of the possible discrepancies in the rating process unveiled by Study 1. The chapter is in three parts. First, although qualitative data were collected to corroborate and explain the rating-related phenomena explored in the previous chapter, the scores awarded by raters on one common writing task will be compared. These data provide insight into rater characteristics which will be compared with the results – if available about the rater – obtained in the first study. Then, the results of the think aloud data collected during the rating process will be discussed. These data will answer the research question which seeks to investigate rater misbehaviour, namely what the sources of the unusual rating patterns in the rating process are.

6.1 A comparison of the scores awarded on a common writing task

6.1.1 The English sample scripts

The two tables below compare the scores on three scripts given by the raters who took part in Study 2. Eleven participants marked English papers and four graded German papers.

Table 21 Scores on the three English writing performances used in Study 2

Code	Criteria	Scores										
		R1	R2	R3	R5	R6	R8	R9	R10	R11	R12	R14
43532	task achievement	5	4	5	4	3+	3	4	5	4	5	(4) 5
	vocabulary	4	3	4	4	4	2	4	5	5	5	4
	style	4	4	4	4	4	2	3	5	5	(5) 4	4
	grammar	4	3	4-5	4	3	3	4	4	4	5	4
	Total	17	14	17-18	16	14	10	15	19	18	(20) 19	17
43537	task achievement	4	3	2	2	3-	2	3	4	2	4	2
	vocabulary	3	3	3	3	2	2	3	4	3	3	2
	style	3	3	2-3	2	3	2	2	3	3	3	(2) 1
	grammar	3	3	3	2	3	3	3	4	3-2	3	2
	Total	13	12	10-11	9	11	9	11	15	10-11	13	7
43573	task achievement	5	5	5	5	5	5	5	5	5	4	4
	vocabulary	5	4	5	4	5	4	5	5	5-4	4	4
	style	5	4	5	4	5	4	4	5	5	3	(4) 3
	grammar	5	4	4	4	4	4	5	4	4	4	3
	Total	20	17	19	17	19	17	19	19	18-19	15	14+
	Sum total	50	43	48	42	44	36	45	53	48	48	38

The first column in Table 21 shows the code of the writing sample, the second displays the analytic writing criteria and the subsequent columns give the scores awarded on the piece of writing by the eleven raters. The analytic scores are summarized for each rater and each script. In the final row of the table, the total scores are presented with a rounding towards the higher number. The scores in the table are highly informative about the idiosyncrasies of the rating processes. First of all, it seems that raters allow scope for flexibility in the final decision-making process: the scores frequently oscillate between two integers. It is also apparent that markers register their hesitations in different ways: some make a record of both scores which might serve as a basis for discussion (3-2), others hesitate, yet give preference to one or the other integer score by putting one figure in brackets: (4) 3. Yet another group indicates by 3 or – indexes that the score is not a clear pass at that level. This marking procedure is a conscious effort on the part of the raters to reach an easier agreement in the final scoring process, as it was also confirmed in the interviews. Apart from formal issues,

the table is also indicative of the individual differences between raters marking the same papers. By simply adding up the scores, the raters can be rank-ordered according to their strictness on the three tasks. For the calculations, as in the usual marking practice, the higher scores will be used as a basis of calculation if there are two scores in a cell. The mean score on the three tasks was 45, those below the mean and thus stricter than average are the following raters: Rater8 appears to be the strictest with 36 points closely followed by Rater14 and 38 scores. Still below the average and thus harsher than the rest is Rater5 with 42 points and Rater2 with 43 points. Very close to the average is Rater6 with 44 points, and Rater9 exhibits a clear middling position with 45 points. Those above the mean and more lenient than the average are Rater3, Rater11 and Rater12 with a total of 48 points. Rater3 and Rater11 show a complete agreement in their total scores, but not according to the analytic scores, though. Rater 12 is slightly different in the distribution of scores. Rater 1 is the most lenient, but her relative leniency is smaller than the relative strictness of the strictest marker, Rater8. These data will be compared with strictness data obtained from Study 1.

Although with such small sample sizes, it is not common to use statistical parameters to characterize the sample, but for the ease of comprehension, descriptive statistics will be used.

Table 22 Descriptive statistics for the total scores of the English writing samples

Statistics		Sample 1	Sample 2	Sample 3
N	Valid	11	11	11
	Missing	0	0	0
Mean		16.18	11.09	17.73
Std. Error of Mean		.85	.66	.57
Median		17.00	11.00	19.00
Mode		14(a)	11	19
Std. Deviation		2.82	2.21	1.90
Variance		7.96	4.89	3.61
Skewness		-.91	-.14	-.90
Std. Error of Skewness		.66	.66	.66
Kurtosis		1.05	.27	-.12
Std. Error of Kurtosis		1.27	1.27	1.27
Range		10	8	6
Minimum		10	7	14
Maximum		20	15	20
Sum		178	122	195

a Multiple modes exist. The smallest value is shown

As Table 22 shows, Sample 43573 was considered the most successful by the raters, with a mean score of 17.73. This is only slightly better than the second best, Sample 43532, with a mean of 16.18. The median is 19 for Sample 43573 and 17 for Sample 43532. It seems that Rater12 and Rater14 were the two exceptions to the general pattern whereby all raters set up the same rank order for the three papers: Sample 43573 was the best paper, Sample 43532 the second most successful and S43537 was the lowest scoring performance in this set of papers. Both the mean and the median provide additional confirmation to this order. The standard deviations present information about the unanimity of the decision: the highest degree of agreement was in

the case of the best paper, with the smallest SD of 1.9. There was a more uneven spread of scores in the case of Sample 43532, with the highest SD of 2.82 and with the largest range of scores (10). The skewness figures describing the symmetry of the distribution are all negative, which suggests that all the performances are above the mean. Kurtosis also provides information about the shape of the distribution. The first two samples have a positive kurtosis value indicating a leptokurtic distribution in which scores are more clustered around the mean. The third sample has a negative kurtosis figure indicating a less peaked distribution. In sum, the descriptive statistics show that the three papers selected were of different performance levels which the raters with two exceptions assessed in a highly similar manner. Rater12 and Rater14 agreed with their colleagues as regards the weakest paper, but they did not share their colleagues' value judgement about the best paper.

6.1.2 The German sample scripts

A much smaller database is available for investigations amongst the German raters. The proportions of raters applied in the study reflect the real proportions in which candidates take English and German examinations in the Foreign Language Examination Centre of the Budapest Business School. In Study 2, four teachers of German took part, and they also marked the same three German writing performances. The results of the scoring are presented in Table 23.

Table 23 Marks for the German scripts in Study 2

Code	Criteria	Scores			
		R4	R7	R13	R15
44322	task achievement	1	2	3	2
	vocabulary	1	2-	2	2
	style	1+	2+	2	2
	grammar	1+	3	3	1-2
	Total	4+	9	10	8
44344	task achievement	1	3	4	4
	vocabulary	2	3-/2	3	3
	style	3-	3	3	4
	grammar	3-	3	3	3
	Total	9-	12	13	14
44411	task achievement	5	5	5	5
	vocabulary	5-	5	4	4
	style	5	5	5	5
	grammar	5	5	5	5
	Total	20-	20	19	19
	Sum total	33	41	42	41

There does not seem to be any difference in registering hesitant rating attitudes between the German and the English markers. Here, again, scores are marked up or down in order to facilitate the final agreement procedure. This table also reveals differences between the raters. As for the strictness of raters, one marker, Rater4 with 33 points is below the mean score of 39. Rater7 and Rater15 are equally slightly above the mean in their strictness, and Rater13 is just one point beyond the previous two markers. All in all, these three latter markers display a highly similar rating behaviour concerning the total score, but as it is apparent from Table 23, on the analytic scale this similarity is less pronounced. The descriptive statistics, although almost meaningless on

such a small sample, adds further details to the general rating profile but focuses more on the writing performance from the perspective of the candidates. For the descriptive statistics the integer and the double scores were rounded up.

Table 24 Descriptive statistics for the German scores

Statistics		Sample 1 S44322	Sample 2 S44344	Sample 3 S44411
N	Valid	4	4	4
	Missing	0	0	0
Mean		7.75	12.00	19.50
Std. Error of Mean		1.315	1.080	.289
Median		8.50	12.50	19.50
Mode		4(a)	9(a)	19(a)
Std. Deviation		2.630	2.160	.577
Variance		6.917	4.667	.333
Skewness		-1.443	-1.190	.000
Std. Error of Skewness		1.014	1.014	1.014
Kurtosis		2.235	1.500	-6.000
Std. Error of Kurtosis		2.619	2.619	2.619
Range		6	5	1
Minimum		4	9	19
Maximum		10	14	20
Sum		31	48	78

a Multiple modes exist. The smallest value is shown

It appears from the mean that three rather different papers were marked in terms task achievement. Sample 44322 was the weakest paper, and Sample 44411 the best. There was a high degree of unanimity between raters concerning the assessment of the best paper, which can be seen from the measures of dispersion (standard deviation, variance and the range.)

As an introduction to the analysis of the qualitative data, the results presented in this chapter so far seem to point to the following conclusion. Both the English and the German markers adopt a similar approach to registering scores by allowing scope for discussions in the final agreement procedure. There is no sign of raters acting as "scoring machines" (Linacre, 2003-6), they all have their individual interpretations of the rating criteria. Nevertheless, these individual interpretations do not result in marked differences in the scores for the majority of the raters. There still seem to be outliers among the raters. In both cases though there appear to be raters behaving slightly differently. The aim of such a study is to identify these raters, monitor their rating practice for an extended period of time and take the necessary steps prompted by the situation.

6.2 Raters' displayed and perceived rating behaviour

The think aloud protocols and the interviews in the second study were expected to highlight important aspects of rater behaviour. Whereas the verbal reports made during the rating process provided data from which conclusions could be drawn indirectly about the rating process and idiosyncratic rating behaviour, the same features were to be confirmed by more explicit information obtained from the interviews.

6.2.1 Rater and rating scale interaction during the rating process

The transcript of the concurrent verbal reports of the rating process yielded data exceeding 30 000 words. Although all scripts have been analysed according to the categories presented in Table 4 and Table 5, and the computer program stores the text chunks arranged in the assigned categories, only those aspects will now be presented which bear direct relevance to the research question. It should be noted, however, that

the data obtained offer an immensely rich source of information on the rating process which will be used for various purposes during the test development process in the future. The discussion of the results will focus on features of the rating process which emerged with the highest frequency in each category. The first group of observations is related to the performance dimension as established by Bachman (1990) and quoted in the theoretical framework in the current study. Observations related to each performance dimension will be summarized, exemplified by data from the tape-scripts and conclusions will be drawn. Although the emphasis is clearly on the qualitative aspect of the study, Table 25 summarizes how frequently raters made references to the most salient features of the rating process.

Table 25 Occurrences of comments made by the raters

Codes and subcodes	**Comments related to the codes and subcodes**	
	No.	%
1.1 Candidate		
1.1.1 Hypothesizing about	8	0.6%
1.1.2 Advice/teaching	20	1.5%
	28	**2.1%**
1.2 Criteria	7	0.5%
1.2.1 Task achievement	47	3.5%
1.2.2 Vocabulary	46	3.5%
1.2.3 Style	47	3.5%
1.2.4 Language use	41	3.1%
1.2.5 CEF levels	3	0.2%
1.2.6 Problems with	17	1.3%
	208	**15.6%**
1.3 Task/prompt	18	1.4%
1.3.1 Relevance	1	0.1%
1.3.2 Instruction	4	0.3%
1.3.3 Problems with	17	1.3%
	40	**3.0%**
1.4 Performance/text		
1.4.1 Length	30	2.3%
1.4.2 Layout	49	3.7%
1.4.3 Lack of/presence of info	59	4.4%
1.4.4 Lifting	23	1.7%
1.4.5 Specific comments	4	0.3%
1.4.6 Overall impression	35	2.6%
1.4.7 Reading or reference to the actual text	152	11.4%
1.4.8 Positive remarks	104	7.8%
1.4.9 Negative remarks	145	10.9%
1.4.10 Neutral remarks, commentary	45	3.4%
1.4.11 Mistakes	145	10.9%
1.4.12 Prefabricated memorized chunks	9	0.7%
	800	**60.1%**
1.5 Rater		
1.5.1 Self	11	0.8%
1.5.2 Rater pair	3	0.2%
1.5.3 Rater group/standardization	4	0.3%
	18	**1.4%**
1.6 Rating process		
1.6.1 Sequence	18	1.4%
1.6.2 Rating process technicalities	58	4.4%
1.6.3 Rhetorical question	31	2.3%
1.6.4 Tech talk	43	3.2%
1.6.5 Comparison	20	1.5%
	170	**12.8%**

1.7 Score	1	0.1%
1.7.1 Analytic	20	1.5%
1.7.2 Total	29	2.2%
1.7.3 Problems with	3	0.2%
1.7.4 Assessment sheet	4	0.3%
1.7.5 Pass mark	11	0.8%
	68	**5.1%**
Total	**1332**	**100.0%**

The most frequent comments concern the written performance itself: mistakes and negative remarks together with citations from the texts account for more than third of the coded segments. It is interesting to note that the rating criteria feature in the segments with almost equal proportions, with a percentage about 3.5 % each. This confirms the finding of Study 1, according to which raters attribute equal attention to each criterion. Additionally, the dominant role of the language use criterion suggested by earlier studies (McNamara, 1996) is not supported by these data, either. The category of language use is the least frequently mentioned criterion among the four, with 3.1%. The frequency of occurrences of the various codes, however, provides little valuable information. It is more intriguing to explore what verbal reactions those figures mask.

6.2.1.1 Performance dimensions: task

During the rating process the task itself is fairly frequently referred to by the raters. Although reading the task prior to the actual marking process and regularly going back to it might appear an inherent part of the rating process, in practice, this is not always the case. From my data, however, it seems that the raters do in fact read the task very carefully before reading the sample performances themselves:

Rater 2

First of all I read … Now I am reading carefully what the task is. Even having read the task, it is sometimes difficult to find my way through…. Now I am reading it to myself, as usual.

Rater 11

Let's see what the task is. The task is … oh yes, I remember this, they have to write a letter to the Chinese tour operator … yes … in 150 and 200 words. And let's see the prompts … what they are expected to include in the letter … yes, that's it. And they also have to attach a … no, they don't actually have to attach … but include in the text. All right then.

Some raters, instead of verbalising their thoughts, while reading the task explain what they are doing.

Rater 12

Silence. I am reading the task now.

Reference to reading the task quite rightly does not appear only at the initial stage of the correction period but also from time to time during the process.

Rater 5

Silence. Sorry, but meanwhile I had to go back to the task itself.

While correcting the letters, markers made frequent references to problems in the task or how the task could be improved. One of the most frequent pieces of criticism was the name and nationality of the addressee which, because of the existing cultural difference

caused unforeseen difficulties for the candidates. This aspect of the task also interfered with the actual scoring.

Rater 12

"Well, these Chinese words, do a bit confuse the reader.

Rater 5

I am not going to consider whether the candidate got the Chinese name correct. My goodness, Dear Mr Zsu, this Chinese name is no good, we shouldn't have had this Chinese name, we only confused the poor candidates.

Rater 5

Now then, Dear Mr Zsu Helling … wow, I have got the third variation."

Apart from making comments on some evident problems with the task, some raters even go as far as to suggest what amendments to the task should be made. This kind of comment clearly testifies that the rater in issue is also involved in test task development and probably has a highly constructive approach to the test validation process.

It is also interesting to note that whereas one rater (Rater8) would include more prompts in the task, the other (Rater2) feels that referring to each existing prompt poses a difficulty to the candidate in terms of text length.

Rater 8

This is a very good task, although in point 3, I would have included prompts eliciting information about accommodation … which students have in fact included in the letter.

Rater 2

I think this is a slightly long task … the tasks requires too much writing, it expects the candidate to write about too many things. Because the task, so the task, well compared to what the task expects, this is a rather meagre letter, even if the number of words exceeded the upper limit. This is so because the task itself expects a bit too much.

These discrepancies can be removed in the pretesting phase of the task or may inform the modification of the task before it is banked. As for raters' attitude to the task during marking, it appears that the rating process is indeed an interaction between the marker and the task as well as the performance. Markers try to adopt an emphatic stance towards the candidate in admitting to the weaknesses that the task may exhibit (Rater5).

Rater 5

Actually, this „itinerary" thing, this might be a problem for some because they send the tourist to either too many places or to too few, and it is difficult to score this.

6.2.1.2 Performance dimensions: performance

While discussing raters' comments related to the writing samples, first the frequencies of the codes need special mention. The most frequent occurrences were related to candidates' mistakes. These data obviously provide useful feedback on candidates' general weaknesses, and can be fed into the teaching process. Positive remarks, namely acknowledging and praising achievement were the next in terms of frequency, followed by comments related to candidates' failing to include information in the piece of writing which was required by the tasks. Among the most frequently specified mistakes lifting, the verbatim repetition of the prompt was the most common. Not surprisingly, as the task involved writing a letter, layout and length were also in the

focus of raters' attention. In terms of mistakes, raters noted problems in relation to each criterion. Grammar, style, discourse features as well as the global task achievement criteria were each touched upon, although not in equal proportion. Next, comments related to each type of criterion will be exemplified. It appears from the number indicating how frequently raters referred to different criteria that *Task achievement* is considered an overarching criterion and as such is less often mentioned. Probably the score given for Task achievement emerges as a result of the other three subscores, or at least is influenced by the criteria more specifically tailored to the assessment of a well defined aspect of writing.

Rater 9

Well, this is a sort of task completed but rather poorly.

Rater 4

… doesn't have a clue about this, doesn't actually have a clue about the whole task.

Raters frequently point out problems related to *Style* including discourse features. Mistakes related to this criterion, however, are not commented on in detail: usually a general positive or negative remark is made without highlighting special problem areas associated with this criterion. Raters generally do not go beyond the actual wording of this criterion, which might indicate that aspects of style are viewed holistically rather than analytically.

Rater 14

Well, the style is definitely not business style.

Rater 4

This is stylistically inappropriate. After all, we are not talking to each other in the street. This was meant to be a business letter.

Rater 1

Total lack of cohesion, also ignored basic rules of text construction.

Vocabulary is also frequently criticized, especially with regard to the specialized vocabulary required by the task type and the content of the letter. Cross linguistic influences are also sometimes mentioned.

Rater 2

"Of course" is a weird structure, and "of course" anyway is out of place here.

Rater 15

The wording here is extremely primitive…

Rater 15

Instead of " seek" s/he uses another verb. Unfortunately. This is a mistake in vocabulary.

The most detailed references are made to *Grammatical* mistakes. It is important to note that not only the nature of the mistake is more specifically described than in the case of other mistakes, but also the seriousness of the mistake is defined. The current analysis uses the word mistake in a general sense, but markers tend to point out the difference between mistakes and errors. They pay attention to such details as to whether the mistake made is a recurrent phenomenon or only an occasional slip.

Rater 12

I underline this a bit… but actually this is not really a mistake.

Rater 2

And now the question arises how strict the rater is. So far s/he has used this structure correctly, so obviously this is just a slip now.

Rater 5

April with a small a. I will underline it, but let's make it a slip only.

Rater 7

It is so badly formulated grammatically that I only have a faint idea of what s/he is trying to say.

It is difficult to decide whether raters attribute truly equal attention to each assessment criterion. Nevertheless, from their comments it becomes apparent that they scrutinize the text carefully and meticulously from each aspect that the rating scale includes, and this eliminates the possibility of giving the same score across all criteria. It also appears that there are criteria in the application of which they are more at ease. Such a criterion is grammar and language use, whereas style is an aspect of writing assessment which markers appear to be less comfortable with. There are no signs as yet of the halo effect, and no criterion is mentioned as more important than the others. Neither is there explicit allusion to the global impression overriding the analytical approach to the assessment.

The high frequency of the utterances in the category which testifies to raters' monitoring the presence and absence of the prompts and the points required by the instruction is highly reassuring from the perspective of the validity of the rating process. It confirms that markers are fully aware of the task requirements, and they are making a continuous effort to adhere to the prescribed guidelines in terms of the task.

Rater 1

S/he is expected to write about four things. Good.

Rater 2

Now let's see the bullet points; which are the ones that s/he has included and which are the ones s/he has omitted. Recommended destinations… that's great. Available types of accommodation, …she provided the whole scale from a to z. A 10 day itinerary would have been long but s/he wrote about so many things that would make up for ten days. Trial costing … well trial costing actually is missing but s/he makes a hint at prices.

Rater 7

There are five points in the task that the candidate has to include in the letter. First of all I will check how many of the points this candidate has covered.

Following the prompts is both a prerequisite for the success of task achievement and a common source of problems at the same time. It creates difficulties both to test-takers and to markers in terms of defining the dividing line between sticking to the prompts and lifting. It is indeed difficult to decide what could still be considered task achievement and what is to be regarded as verbatim repetition of the prompt. It is evident from the markers' comments that they are sensitive to the issue of lifting. Nevertheless, what is regarded as lifting apart from the obvious copying of the prompt is subject to raters' individual interpretation.

Rater 3

It will be quite easy to write this letter. Never mind, I will have to make sure that there shouldn't be any lifting in the text.

Rater 3

And now comes that s/he has to give a sample itinerary and give a couple of recommendations. Well, the first two sentences are perfectly identical with the prompts. So far s/he hasn't done anything.

Rater 6

… now around the middle I am beginning to feel that s/he is following the prompts too closely … well, of course much of the vocabulary comes from the prompts.

Rater 7

Of course s/he doesn't make any mistakes here because s/he has lifted this whole passage from the prompts.

Similarly, markers are equally strongly sensitized to memorized chunks of language, which, in the case of a letter writing task can be represented by completely memorized formulae and even complete memorized letters.

Rater 5

These are lovely memorized coursebookish sentences. It is quite difficult to evaluate these memorized sentences, I have to say. Because by putting down these memorized sentences s/he torpedoes my efforts at the proper assessment of vocabulary and grammar.

Rater 5

Gotcha! I remember reading lovely sentences like these in a book called New Guide.

Rater 9

It is easy to see from this letter that s/he has memorized the parts of the letter but manages to use the memorized things quite cleverly.

What emerges from these excerpts is that the assessment of sentences, phrases which have been either memorized or taken from the prompts and inserted in the text without properly embedding them in the context is an issue for serious consideration. Firstly, it should be given critical thought how a task offering explicitly the opportunity of lifting can be altered. Additionally, raters' observed sensitivity to the use of ready-made chunks of language suggests that this issue should be fully discussed during the rater training sessions as well as in the standardization meetings. Consensus should be reached as to how these undesirable candidate practices should be dealt with. Furthermore, this information should definitely be channelled back into the teaching process.

6.2.1.3 Performance dimensions: candidate

In the rating process the candidate also plays an important role. Although they are only indirectly present through the performance, raters link the inanimate product to its producer and make references to them. The two most common remarks concerning the candidate are in the form of advice on what s/he should have done or hypothesizing about the candidates' characteristics related to the completion to the task.

Rater 15

It is clear that s/he has no practical background knowledge. S/he has no idea of what to sell, how to sell it, but this is another issue now.

In these comments, the specific aspect of the language examination is manifested. There are frequent references to the field of specialization, nevertheless raters are capable of remaining in their roles and act as language teachers rather than special subject instructors (Rater 5).

Rater 5

This is indeed a good itinerary. Obviously s/he wasn't very good at tourism geography either. Never mind, we don't have to evaluate his/her knowledge of tourism geography.

In addition, cross linguistic influences are disclosed (Rater 15), but apparently not unduly penalized.

Rater 15

S/he must be good at English. (*Remark made by German rater*.) Puts a semi colon after the closing greeting.

Raters' sympathy and a general positive attitude towards the candidate continue to be evident (Rater 5).

Rater 5

Oh poor thing, s/he got probably tired by the end.

Raters might get annoyed by mistakes occurring in spite of drawing students' attention to them while preparing them for the exam (Rater15).

Rater 15

And anyway … what we always draw their attention to is that they shouldn't push themselves in the foreground … I don't see too much of that here. Of course it all depends on how much emphasis his/her teacher laid on this issue.

6.2.1.4 Performance dimensions: rater

Among the performance dimensions, explicit reference to self, that is, to the rater is fairly common. In the think aloud procedures raters seemed to admit openly to

their weaknesses, and they were not reluctant to disclose language problems and difficulties posed by the task itself which they kept struggling with.

Rater 12
Let's leave it as it is. I don't know that, either.

Rater 13
I always seem to forget whether there should be a full stop after the abbreviated form of Limited Company, but rumours say there should be one…

Rater 13
I am not absolutely positive that this is a mistake.

Quite frequent are the allusions to the rater pair. This suggests that during the rating process, the existence of double scoring and the agreement procedure with the rater pair serve as a constant control and invisible monitor for the rater.

Rater 1
I am not quite sure that I can decide this on my own. Luckily there is another rater, too.

Rater 2
I will give 5 for task achievement, which might be challenged by my rater pair. Still, I feel that everything required has been included in this letter.

It seems from the raters' comments that both the rater pair and the agreement procedure are an inherent part of the rating process. The standardization of marking prior to the start of the live marking is also an event which markers heavily rely on and regard as a safeguard of reliable marking.

Rater 2

Yes, now if I think of the 14 points this candidate received, I think that during the agreement procedure with the others I am not really going to insist. If the consensus moves towards 15, I will have no objections but give in because this is a fairly good letter. So here I feel I will give in. I wouldn't approve of anything lower though because 12 is the pass mark and this is an acceptable letter. So it's sort of 14-15 points. No, I think 14 will do.

Rater 2

It is always very important … it really counts when we compare the scores given for the first five letters because then we can discuss problems more extensively.

The standardization of marking has not always been an inherent part of the marking procedure. Whereas general rater training has indeed constituted an important component of obtaining an accredited status for raters, the standardization of marking was introduced some 3 years ago and has become a regular practice since then. In a situation where several markers have to mark the same type of task, standardization creates a common frame of reference which raters heavily rely on (Rater2).

6.2.1.5 Performance dimensions: score

Raters' behaviour during the actual scoring procedure shares some common characteristics. It appears from the data that even during the analytical rating process, markers heavily rely on the impression that the total scores convey. After the meticulous use of the analytical scales, they are ready to go back and modify their originally awarded scores because the total score added up from the analytical scores is either too high or too low in comparison with what they would give intuitively and globally. It should be noted that in the majority of the cases, raters find it necessary to go down with the points rather than to go up. Total scores are frequently interpreted and

calculated with reference to the cut off score of 60 %. This is all the more interesting as in the subtest it is not necessary for the candidate to achieve 60%, only a global achievement of 60% is required to pass the whole test. Raters still have a theoretical cut off point in their minds while deciding on the final score. This theoretical cut off point subconsciously guides their assessment practice.

Rater2

This is a narrow pass. I always tend to bear in mind the pass mark. Yes, it is exactly 12 points.

Rater 3

Well this is a sort of … borderline case for me, a just pass. Something like 12 or 13 points.

Let' see then.. well yes, 3,6,9,11,12. Yes exactly a borderline case.

Rater 7

It is 12 points altogether. But I have to note that personally I don't consider this letter an intermediate performance. So instead of 3 minus for vocabulary I should have given him/her a 2 because this is not intermediate level. If I stick strictly to the descriptor scales the total score will be 12 … but no. This is not intermediate level. I wouldn't give an intermediate exam for this performance. But I have to add that vocabulary may be 2 points eventually.

The think aloud protocols also give evidence of how decisive or indecisive raters are during the rating process. Some seem to be hesitant all through the process and tend to modify their scores. Others, on the other hand are highly decisive and do not contemplate whether the total score is consistent with their intuitive feelings. In addition, locating the candidate's competence in either the upper or the lower realm of the band descriptor is a common practice. This highly individual attitude to decision making was also reflected in the scoring sheets used in the first part of Study 2.

Rater 5

What I said was 4-3-4-4. How much is that altogether? 15 points. 5 points but I think this is a better letter than that, so in retrospect … instead of three for vocabulary I would give him/her a 4. Four times four, which makes 16. I don't like giving four times four because it looks as if I didn't give enough thought to things.

Rater 5

If I follow the assessment criteria, this is 3. But that would make 11 altogether for this letter which I would consider a bit too much. So then now in retrospect I will have to think it over where I could possible deduct some points.

Rater 5

This is 18 points altogether which I find slightly too high. Let me see where I can take off points. For example this one needs to be penalized for including the first name when he addressed the addressee like Mr Zsu Helling, provided Zsu is the first name. Well, there were indeed mistakes here which shouldn't be in a letter with 18 points so I would give this a total of 16.

Rater 9

12 and 3 totals 15, a bit too high but never mind.

When raters feel the need to decrease the total score, quite interestingly, it is the vocabulary criterion where most of them seem to adopt a stricter approach.

Rater 11

I think I should reconsider vocabulary … or at least I should cut one point off … 18-19 … like this.

Rater 13

It might be possible to give 5 points for vocabulary, but I don't really think I should. Looking at it, with all my red marks and all that, well, I don't think this letter deserves 20 points. And where I can deduct points is vocabulary. Of course, with a full awareness of the assessment criteria.

Rater 11

Right then, task achievement is 4, vocabulary perhaps 5, language something like 4. And now style. What on earth is that? Disproportionateness of organization. It is somewhere halfway. It is not that bad … it could be… what shall I give him/her? I like it, this is 5, and where shall we take one off? This wasn't very good after all, vocabulary, yes, there I will take one off. No I won't.

Hesitations and score modifications are always accompanied by a second reading of the text, or at least a quick scan through the text, so even the modifications are mostly well-informed decisions. It appears that even the individually awarded scores emerge as a result of several readings and a careful weighting of various factors.

6.2.1.6 Performance dimensions: rating criteria

The most intriguing area from the perspective of the focus of the research is raters' attitude to the rating criteria. Although they have already been indirectly referred to in the discussion of the scoring procedure, next raters' references and attitudes to the criteria will be discussed in more detail.

Raters consider the systematic use of the rating criteria an essential element of the rating procedure, and even after a long marking history and a thorough familiarity with the assessment criteria, they seem to find it important to pay constant attention to the descriptor scales of the assessment criteria. This might sound a commonplace but anecdotal evidence suggests that novice raters especially who are otherwise experienced

teachers try to make attempts to use their routine obtained through long years of teaching to embark on an assessor career with an examination board. These people do not feel the need to consult the assessment scale, and they refer to their extensive experience and reliable intuitions in being able to decide what constitutes mark 5 or mark 4. Evidence from the think aloud protocols supports that the raters in this study closely follow the assessment scales and heavily rely on what is defined by the criteria all through the rating process. They make continuous attempts to match the features of the letter they are marking with the descriptor bands of the assessment scale.

Rater 2

I always put this sheet in front of me, that is I keep an eye on it all the time, even if, obviously to a lesser degree, but I have it in front of me even after the 200th letter. And now I am going to read it, read the scales. Or at least I run through them like I am doing it now. Now then, task achievement, vocabulary, style and language use.

Rater 5

Well then, the rating criteria. Did I receive any of those? Yes, I did. First I am going to read them, and start with task achievement.

Rater 9

And I had a look at the assessment criteria.

Rater 3

Let's see the rating criteria now: task achievement: appropriate text type including all points. Yes, they have been all included, so it is 5.

6.2.2.1 Rating criteria: Task achievement

This criterion focuses on the appropriacy of the text type produced as well as on the inclusion of all predefined points in the output letter. It is also required by this criterion that the necessary points should be included logically and in appropriate detail. What seems to cause problems for the raters is to make a decision if some of the requirements are met, whereas others are not, within one descriptor band. It is with the application of this criterion that an earlier discussed phenomenon, namely lifting is addressed. Drawing the dividing line between including the bare prompts in the text and including them in the text in an appropriate fashion but without unnecessary details does indeed create difficulties for the assessors. In the following excerpts, direct quotations, exact words from the assessment scale are printed in italics.

Rater 9

Now let's see task achievement. Yes the points are covered but not really logically and neither in enough detail. The necessary points are included – this is 4 points. Point 3 says that there are also unnecessary points included. No, I cannot say this, and I cannot really claim that it is illogical and scanty, so it deserves 4 points.

Rater 10

The *candidate prepares the type of text specified in the task* … all the points are included … they look quite good … now the problem is that the points are usually lifted which is rather annoying.

Rater 11

Task achievement: well *then the candidate prepares a text type specified in the task*... well this is a bit fuzzy… but I think eventually it can be accepted. Let's see again in detail … didn't ask what type of information they require … and did not write about … but it's OK, perfectly OK. This can be around 5 points. For task achievement that is.

Rater 12

This is a dilemma, because I would like to give him/her 4 points according to the scales but two pieces of information are missing so I am inclined to give 3, yes, 3 points at least according to the scales.

Rater 2

Task achievement … I am in two minds between 3 and 4. The necessary information is included in the text … I will give 3 because there were missing bits and the letter does not really match the required text type. *Most of the required information is missing* … Yes this is definitely 3 points, *required information is mostly included*.

Rater 4

Task achievement: *repeats, reproduces the given prompts in fragments*, yes and the whole thing is a mess.

The Task achievement criterion raises a further problem as well. The intriguing question of what exactly is meant by language for specific purposes is manifested in connection with this criterion. Raters are faced with the task of having to act as subject specialists to a certain degree. Markers have to assess how successfully the task was completed, and this obviously involves having to resort to subject specific knowledge as well. The ever-existing dilemma about the extent to which a language teacher is capable of and entitled to assess subject specific knowledge is also a matter of concern for markers.

Rater 5

Let's see: task achievement. Task achievement is much better that with the previous two, because s/he writes about Budapest and what there is. S/he also describes what is good about Lake Balaton and Hortobágy. Actually Hortobágy

is identified with the great Hungarian Plain but let's not be perfectionists, he will be failed for this in travel geography if s/he is a tourism student. So I would really like to give 5 points for task achievement, so let's have a look at the assessment scale. …Logically, in appropriate detail, yes I am happy to give five points for task achievement.

As it has been seen before, raters do not seek to take the role of subject specialist teachers, and they remain conscious of their roles as language teachers. From the perspective of my research the most valuable data are related to possible problems with the rating scale. Raters comment on the difficulty of applying the issue of including unnecessary detail in the text. This is included in the descriptor for point 3 in the task achievement criterion. One rater (Rater8) simply expresses her negative attitude to this, whereas another one (Rater7) also justifies why she finds it difficult to use this in the rating process or why she finds that this is not a well-functioning sentence in the assessment scale.

Rater 8

The fact that *unnecessary points are included in the text* is not very good point, I don't know what to do with it.

Rater 7

This *unnecessary points* issue always gives me a headache. I would think that there is no candidate who would include *unnecessary points* in the letter. I don't know about the colleagues teaching English, but all through my career I have encountered only once … actually this is not quite true but to be more precise I have very rarely encountered candidates who included unnecessary points in the letter. And this makes assessment slightly more difficult. Because this performance would perfectly match band 3 in the assessment scale up to the point of unnecessary details. If it were not for unnecessary details, the letter would perfectly match the band. I don't think the problem here comes from

including unnecessary points in the text but rather from the omission of some of the necessary ones.

Deciding on the score is rendered difficult by applying all conditions within the descriptor. Raters have to decide whether the performance is up to a certain standard and deserves a certain score even if not all the criteria are met that are included in the descriptor band.

Rater 9

Task achievement. There were quite a lot of problems with this task achievement so in the end 3 will be the most appropriate score. Although it is not true that *there are unnecessary points* in the text, but the rest matches the description, namely that the *text is appropriate to the task type and all the necessary information is included*, that's why it is 3 points.

6.2.2.2 Rating criteria: Vocabulary

The Vocabulary criterion seems to be perhaps the most straightforward and the easiest to apply for raters as the results of Study 1 also suggest. It is with this criterion that they are the fastest to make decisions on the scores, and they do not have to ponder about their decisions at length. Only one issue emerged as slightly problematic: the phrase "appropriate to the text". The relativity implied in the phrase does not always give clear guidance to the raters. The issue of including exact words and chunks of text from the prompt continues to present problems. Lifting from the prompt and its consequences are indirectly mapped on the assessment scales, but the problem itself can be ramified by validating the task type and the accompanying instructions. Similarly, memorized chunks of language which stick out of the text are often commented on by raters with a negative overtone.

Rater 1

Vocabulary. Very cautious, what is s/he saying? I would think it is 4 points, mostly appropriate vocabulary, of course words which are also included in the prompt but it is not really noticeable …yes I would give 4 points for this.

Rater 2

As for vocabulary … 3, as s/he does not always use the appropriate words and terminology. It is interesting though because s/he has learnt a lot of terms and phrases but probably his/her general language proficiency is not at the level to apply these phrases properly, to link them neatly together.

Rater 3

Vocabulary: I like it, I'll give him/her 5 points because good terms are included in the text, like destination, facility … Hang on, destination is actually in the prompt. But potential guests, sightseeing, popular destination, opportunity for recreation are all good terms and these have not been included in the prompt. Good then, it deserves 5 points.

Rater 7

Band 2 says that *the candidate almost never uses appropriate terms*. This is in fact true as all the terms in the letter come from the prompt. And those which do not come from the prompt are rather fuzzy. Yet I cannot give 1 point only because that would mean that the use of *words is misleading in many places* or *terms were used inconsistently*. To be honest, if I consider the last sentence this description is perfectly in place but on the whole I feel that this is 2 minus.

While assessing vocabulary, two raters also referred to the existence of a relationship between the assessment criteria, the fact that certain mistakes might belong to more than one criterion.

Rater 5

Vocabulary. Now here again the problem is that the memorized sentences have been put down correctly but here were some minor mistakes in the sentences s/he constructed. Oops, I am not supposed to deal with this at the moment but only with vocabulary.

Rater 9

Vocabulary, well, s/he does not always use the appropriate words or phrases, a couple of polite formulae are missing and a couple of linkers. Obviously this is partly style but partly also lack of proper vocabulary. Vocabulary is 3 points.

6.2.2.3 Assessment criteria: Style

This criterion deals with the discourse features of the text and includes such text features as coherence and cohesion, style and layout appropriate to the genre and the required text type. The think aloud tape-scripts again confirm the need for raters to carefully establish the balance between the sub-criteria met and unmet within one descriptor band, and thus make the final decision and award a score.

Rater 7

Let's see the criteria now… I don't even deal with the band descriptor for 4 points. 3 – *the candidate prepares a text which is acceptable from the perspective of the required text but its style is uneven and there are organizational mistakes in the text. 2 – there are only scant traces of the typical features of the genre.* Well, this would be a bit of an exaggeration so I will give 3 points for this.

This frequent hesitation between two bands is highly characteristic rater behaviour throughout the rating process, and a natural one. It is less common that a rater can make a decision on the scores straight away without any hesitation. In addition, the use of global impressions coupled with a profound knowledge of the scales is fairly apparent in the use of this assessment criterion: before deciding on the score to be awarded, raters usually do not read all six bands but start at a certain point of the scale, usually at the middle, and from there move upwards or downwards.

Rater 13

Style. Again, I will take 3 as a starting point. *The candidate prepares a text which is acceptable from the perspective of the required text but its style is uneven and there are organizational mistakes in the text. There are only scant traces of the typical features of the genre.* Well I don't really know … The candidate prepares an acceptable text, uneven, organizational mistakes … well style… well, features, characteristic of the genre in fact exist in traces, so I would rather go for 2 points.

Rater 3

Style … I would go a bit lower here because I feel that the style is rather informal with all those „I"s. S/he could have used a passive structure or something every now and then. "Be so kind and let me know" is also kind of patronizing. On the other hand there are no organizational mistakes, so I would stick to 4 points in the end.

Finally, it should be mentioned that this is the criterion which is perhaps the least evident for one of the markers. "Style? What on earth is that?", as Rater 11 formulates her problem with the criterion once, and later, "What the hell is style again?" (Rater 11). Here again it seems that style is a criterion the interpretation and a full comprehension of which might pose problems to the raters.

6.2.2.4 Assessment criteria: Language use

Not surprisingly, the issue of lifting is again a frequent source of criticism. Markers note that the lack of mistakes in certain parts of the letter is due to the candidate's memorizing certain formulae and chunks of language, and this is contrasted with mistakes made in parts of the text which the candidate himself or herself made up. It seems that lifting, or word for word quotation of the prompt has such an overarching effect that is penalized three times, in three different assessment criteria. Another major

controversy arises from the interpretation of the number of mistakes. What is defined by the assessment scale allows two alternatives: higher score should be given for more complex sentences and fewer mistakes, and lower scores should be awarded for simple structures and a large number of mistakes. This seems to be a logical stepwise progression but it is in fact purely theoretical and goes contrary to what happens in actual practice. Some of the raters also seem to be very ungenerous with the highest score and this is most manifested with the language use criterion. For some, the existence of even one mistake in language use excludes the possibility of awarding the highest score, for others however, one or two minor mistakes are still acceptable for a total score. Whichever group markers belong to, this is an issue for consideration.

Rater 1

Grammar: logical connection. Does not have too many connections but grammatical mistakes are few. Seemingly is unhappy but does that really matter? Let's see, about seven mistakes … what does *insignificant/negligible number of mistakes* mean? Does negligible mean zero? The difference between insignificant and more? Hang on, *recurring mistakes* … no definitely no repeated mistakes, well, somewhere between 4 and 5.

Rater 2

Language use, spelling. *Language use, spelling, mostly simple structures, grammatical mistakes and repeated grammatical mistakes … without distorting the meaning*. By no means can I say that this is 4, that this performance is characterized by well-formed sentences, adequate structures. Where the structures are adequate, that is a memorized chunk of language which should of course be appreciated because it is important in the case of a business letter, but the structures which the candidate himself/herself created are not adequate. So I would say this would be more like descriptor band 3.

The memorized sentences are in fact a major source of raters' strictness as it is apparent from the comment made by Rater 2. From the frequent mention of the problem they might cause, it seems that raters are highly sensitized to this form of candidate behaviour.

Rater 5

Language use and spelling. Now I have a problem here again. These memorized sentences have been perfectly put down, but on the other hand s/he makes such major mistakes as "a lot of wine region" and "a lot of castle" in singular. This would be pre-intermediate level. And it is also difficult to evaluate the memorized sentences, I have to admit. And this "I thank you" also gave me a bit of a shock at the beginning. So *badly constructed, fragmentary sentences, a number of mistakes* or rather 3: *mainly simple sentences, grammatical mistakes*. If I follow the scoring guide, then it is 3.

Rater 6

Language use and spelling. Well, there aren't too many mistakes in the text. In spite of that, I wouldn't really like to give 4 points because s/he follows the prompts too closely and uses quite a lot of phrases from them. I can't see much of the candidate's own work. I will give 3 points with the remark that my global impression is that this is somewhere in the middle.

Rater 6

I think that even these memorized sentences are far too simple. S/he does not use any complex structures. Does not take a risk. And I feel that the sentences in the middle about thermal baths and Tokaj are also ready made sentences to a certain extent. The sentences are extremely simple. Even so, s/he makes constant mistakes with singular and plural, so it is extremely difficult. It is extremely difficult because probably this candidate is not very good grammatically. But for all that I cannot say that the sentences are fragmentary, what I can say is that the text consists of mainly simple sentences … I choose this category and give 3 points for grammar.

The previous two comments made by Rater 2 also confirm that there is a conflict between grammatical accuracy and memorized chunks of language.

Although the different criteria within the analytical scale focus on well identifiable aspects of the skill, they show certain inherent interrelatedness, however undesirable the cross-contamination of descriptor bands might be. There are cases when the observed phenomenon itself extends beyond the boundaries of one criterion, and the criteria cannot be used in isolation. The following excerpt indicates that although the language use criterion focuses chiefly on grammar, it is also linked to style as it is difficult to evaluate sentences in isolation even if the centre of attention is on language use.

Rater 10

> And the last criterion: *language use and spelling*. Well, quite often the plural s has been omitted. It is not true though that the text consists of mainly simple structures because there are quite a few obvious attempts at more complex structures ... *grammatical mistakes or recurrent mistakes without distorting the meaning* ...actually, the meaning hasn't been distorted but you don't have to be perfect for 4 points. Because the *text consists of well-constructed sentences, adequate grammatical structures, but there are several grammatical and spelling mistakes*. Let's give him/her 4 points, but there are a couple of simple sentences which could have been linked to the previous ones and then it would have sounded much better.

The hesitation between two band descriptors and consequently two consecutive scores continues to be prevalent with the language use and spelling criterion. There seems to be only one rater who does not give excessive consideration to what different stipulations within one descriptor band might entail but makes straightforward and

prompt decisions, sometimes based on one selected, probably dominant point in the scale.

Rater 8

Grammar and spelling: in spite of the mistakes I can't say that the mistakes distort the meaning, so I would give 3 points here.

It seems that there are certain recurring elements that characterize raters in the use of all rating criteria. On the positive side, it should be noted that raters heavily rely on the rating criteria and follow them closely throughout the marking process. This is confirmed by the frequent word for word mention of the descriptor bands in the think aloud protocols. A careful investigation and a need for an accurate interpretation is indicated by the indecisiveness raters show before deciding on the final scores. On the negative side, however, there are also issues which require attention. The exact meaning of certain quantifiers, such as "few", "negligible", "a number of" and similar phrases should be made explicit during the standardization of the rating. In addition, what needs further consideration is to decide what proportion of the assumptions in the descriptor should be met in order to get the given score.

6.2.3 Comparison

A fundamental difference between norm referencing and criterion referencing lies between what candidates' performances in these two different forms of assessment are compared to. Whereas norm referencing implies the comparison of performances directly with each other, criterion referencing means comparing performances to an external yardstick, the assessment scale. The examination which is in the centre of this

study is criterion referenced, and therefore subjectively scored performances are directly compared to the assessment criteria specifically developed for each task and evaluated on the basis of the scales. Consequently, raters are expected to compare performances to the standards laid down in the assessment scale rather than to each other. The empirical data, however, indicate that the comparison of performances is not an uncommon rater practice. It appears that it is inevitable to be influenced by other performances during the rating process. At the same time, it is important to note that these comparisons are not the primary sources of raters' judgements: they are chiefly applied to fine-tune the given scores or to justify a decision eventually made. Raters make references to other performances in other contexts as well. Raters' comments suggest that they need a certain amount of time and a certain number of scripts to mark to feel comfortable about being fair with the awarded scores.

Rater 2

With the first letters, and not because I am doing this for you Eszter, but with the first letters I always go through them for a second time. I am doing this because it is worth the effort, it will pay off later on. In my experience if I am not careful enough and do things in a rush at the initial stage, then I will be hesitant and indecisive all through the rating session. So the first couple of letters should serve as yardsticks.

Rater 4

I may be a bit too harsh, but this is the first script.

Rater 1

Having read the first script for the first time it seemed much better than after reading the third script.

…

> I have the feeling that I might give slightly different marks at the end of a rating period from what I have given now. I feel now as if I were at the beginning of the rating session, in a sort of standardization session when we tune in and get used to the assessment criteria. So I suppose that this is not yet the real thing.

Although comparing performances might seem an out of place practice in criterion referenced assessment, creating the context for the assessment of the specific task entails a specific form of comparison. This familiarization and initial adjustment probably cannot happen without actual comparisons, but care should be taken that the comparisons of performances should be primarily informed by the assessment scale.

6.2.4 Common European Framework of Reference

A useful finding of the study showed that at the time of data collection when the examination centre was only in a transition period of linking the examination levels to those of the Common European Framework of Reference, and the assessment scales were not yet aligned with the new levels, some raters made clear and explicit references to the new system of values. This indicates that during the intensive familiarization work the new levels have been ingrained to such an extent that there are good chances that their use can be incorporated into the evaluation procedure smoothly, unnoticeably, and without major problems.

> **Rater 1**
> This would even do for a C1, s/he rounded it up so beautifully.

> **Rater 1**
> At B2 level they should be able to write an acceptable letter.

Rater 9

Well, this task has not been completed at B2 level. This is more like B1, so this does not come up to the pass level.

Rater 15

And anyway, if we have to adjust our levels and go down slightly to B2, then this performance perfectly matches that level. It does not extend into C1 that much.

This final statement makes an explicit reference to the harmonization work which entails that the originally set intermediate level which is B2 and extends into C1 should be adjusted so as to represent a clear level, namely B2.

6.2.5 Conclusion

In this chapter the most important rater characteristics have been identified and illustrated with excerpts from the think aloud protocol tape-scripts. Although altogether 53 categories were used for the coding process, many of the issues are not discussed now as they bear no direct relevance to the research questions. Based on the frequency of codes and the relevance of the comments, the results of the analysis of the think aloud procedures can be summarized in the following points.

General assumptions related to the rating process

1. The rating process yields important information about both the instructions and the prompt, and suggests modifications which should be implemented before a task is banked. This claim is consistent with Shaw's (2004a) discussion of the task-related issues that need special attention in the validation of the IELT Writing test.

2. The standardization of marking is an essential prerequisite of fair and reliable marking. This result is similar to the finding of Lumley (2005), who likewise concluded that although general rater training is necessary, but each operational rating session should be preceded by a reorientation focusing on the actual task.
3. Heavy reliance on the second marker's views and the agreement procedure is an inherent part of the rating procedure. This is an issue that needs further investigation: whereas individual raters' rating processes have been extensively researched, much less attention has been devoted to how raters in the double marking procedure come to an agreement concerning the final, agreed scores.

Rating process characteristics

4. The careful use of the assessment scale and a frequent reference to it guides the rating process.
5. The final score emerges as a result of an adequate consideration of all analytic assessment criteria. These latter findings, however, are based on observations made in experimental circumstances. As Lumley (2005) has also pointed out, this careful attention to all criteria might only reflect an idealized rating process staged specifically for the think aloud procedure. It can only be hypothesized that this would also be the case in an operational rating session.
6. Raters are specifically sensitized to features such as the verbatim repetition of the prompt and the use of memorized chunks of language. It appears that it is not easy to draw a borderline between verbatim repetition and paraphrasing the prompts, as Lee and Kantor (2003) have also found in their study. This is clearly an issue to consider and clarify in the rater standardization session. Additionally, the band descriptors should give explicit guidance as to how to deal with the use

of memorized phrases and prefabricated lexical chunks, as Shaw (2004a) also suggests.

7. Most raters have their idiosyncratic sources of deviations which they are specifically sensitized to, for instance to the overuse of "I" in a business letter. As Vaughan (1991) also argues, "raters are not a tabula rasa, and do not, like computers, internalize a predetermined grid that they apply uniformly to every essay" (p.120). The salient features in a writing task that might strongly influence the raters in her study include handwriting, text length and the "unique use of an extended metaphor" (p.121).

Rater and rating scale interaction

8. In the use of the assessment criteria, shortcomings of various aspects of the rating scale might become evident: for example lack of consensus regarding the exact meaning of quantifiers. This has been proposed by numerous writers (e.g. Davidson, 1991; Hawkey & Barker, 2004; Shaw, 2004a) and confirms North's (2000) claim according to which definiteness, clarity, brevity and independence should be essential features of the descriptors.
9. Not all assessment criteria are treated in an identical way: the task achievement criterion serves as an overarching criterion and is considered as having a strong link to the other criteria. Numerous studies (e.g. Eckes, 2005; Engelhard, 1994; Kondo-Brown, 2002; Weigle, 1994; Wolfe, 2004) have similarly concluded that raters all have their idiosyncratic interpretations of the rating scales. The results concerning the criterion playing a dominant role, however, seem to be contradictory.

10. Global impressions and scores emerging as a result of analytic scoring are not always in complete agreement: this inconsistency may result in post-hoc score modification. The importance of overall, intuitive impression in the scoring process was also reported by Lumley (2005), who found instances of raters' reconsidering previously given scores together with a tendency to award the same scores across all rating categories.

Chapter 7: Perceived rater behaviour

Introduction

To obtain further details about the rating process and how raters explicitly see their roles in the assessment procedure, the interviews with the fifteen raters participating in Study 2 were analysed. The complete tape-script exceeded 53, 400 words and as with the think aloud tape-scripts, only the heeded aspects of the interviews will be discussed in detail. The interviews focused explicitly on perceived rater behaviour and sought to disclose the aspects of rating behaviour which might divert raters from the assessment scale. The chief categories were based on identified rater misbehaviours (Linacre, 2003-6) and extended by further typical distinctive traits. Thus, the following rater characteristics were investigated in detail: leniency and harshness, extremism and central tendency, the halo effect, the application of response set and the "playing it safe" strategy. Each form of rater misbehaviour will be discussed and exemplified in the chapter. First, however, some initial assumptions will be made on the basis of quantifying the interview data. Table 26 indicates how frequently the coded segments related to rater characteristics occurred in the interview transcripts.

Table 26 Frequency of comments in the raters' interviews

Codes and subcodes	Comments related to the codes and subcodes	
	No.	%
1 Leniency, severity	17	2.8%
1.1 Leniency, generosity	26	4.3%
1.2 Comparison with rater pair	22	3.6%
1.3 Agreement procedure	33	5.4%
1.4 Global and analytic rating	19	3.1%
1.5 Adjustment to rater pair	28	4.6%
	145	**23.8%**
2 Extremism, central tendency	15	2.5%
2.1 Zero/maximum score	10	1.6%
2.2 Reasons for zero/maximum	5	0.8%
2.2.1 Zero	49	8.1%
2.2.2. Maximum	43	7.1%
2.3 More frequent	11	1.8%
	133	**21.9%**
3 Halo effect	7	1.2%
3.1 Most important criterion	23	3.8%
3.2 Least important criterion	15	2.5%
3.3 Relationship between criteria	50	8.2%
	95	**15.6%**
4 Response sets	**4**	**0.7%**
5 Playing it safe	**5**	**0.8%**
6 Instability	35	5.8%
6.1 Factors influencing assessment	66	10.9%
6.2 Factors resulting in strictness	45	7.4%
6.3 Factors resulting in leniency	20	3.3%
6.4 Within rating period consistency	22	3.6%
6.5 Across rating period consistency	13	2.1%
6.6 Blackout	25	4.1%
	226	**37.2%**
Total	**608**	**100.0%**

Although the data presented in Table 26 allow making only tentative judgements, it appears that inconsistencies were the most frequently mentioned forms of rater misbehaviour. It is also interesting to see that raters were more willing to comment on their strictness than to give reasons for or justifying their leniency. In discussing leniency and strictness, raters fairly often referred to the agreement procedure. This fact seems to support an assumption made earlier according to which the raters' agreement procedure might also serve as a rich source of information and deserves further investigation. Rater characteristics that can be rightly termed aberrant (Wolfe, Moulder, Bradley & Myford, 1999), namely response sets and playing it safe as rating strategies did not appear to be typical features of the raters interviewed.

7.1 Rater leniency and harshness

One of the most frequently researched areas of rater behaviour is rater leniency and harshness. The present investigation focuses on two issues related to leniency and harshness: what is the extent to which raters are aware of their own generosity or strictness in awarding scores, and how do they relate to this, or in other words, what justification do they give to their supposed deviation from the norm. Five sub-codes were created within this category: apart from discussing harshness in general, this feature was rather frequently mentioned in relation to the comparison with the rater pair. The role of the rater pair was also discussed in terms of the agreement procedure and the possible adjustment to the rater pair's supposed leniency and harshness. Finally, leniency and harshness were also brought up with reference to global and analytic rating.

7.1.1 Leniency and harshness in general

Although the majority of raters assigned themselves to the ’neither too lenient, nor too harsh category’, it appeared that some of them indeed were able to assess their level of strictness.

Rater 1

Middling. Neither too lenient, nor too strict.

Rater 2

The question was whether I consider myself strict or lenient. I haven’t answered your question. I would like to consider myself neither, I would like to see myself as someone being close to the fair average.

Rater 4

I am somewhere in the middle. Neither strict, nor lenient. … But maybe I am a bit harsher than the others.

Rater 3

I don’t think I belong to either extreme, the strict group or the lenient one. But if I had to decide I would say I am slightly stricter than the average. But I don’t think I am significantly stricter than the others.

An interesting finding is that according to some of the interviewees, harshness and leniency are not a permanent trait but a relative feature. There are factors which may influence raters’ level of strictness.

Rater 5

Strictness changes. It depends on my mood. But on the whole I think I am more on the lenient side.

Rater 11

I think I am somewhere in the middle. I am neither too strict, nor too lenient. But I also think that sometimes I am sort of fluctuating. When I am in the middle of a batch of papers to be marked, my judgement changes.

For one rater, a purposefully heightened level of strictness may have pedagogical functions, and it is also related to the purpose of the assessment, the purpose of the test.

Rater 1

Everything is relative there (*in classroom assessment*). It always depends on the purpose of the test and on my purpose with the test. And I correct the test and students ask me about the mark. It can be any mark I want to assign to the test. It depends where I put the cut score, how I define the measurement units. If it is a lazy lot, I can say that the cut score for mark 5 is 90% and things like that. Or alternatively I can say that I am glad they did it, and I mark them up, so on the whole I change the level of strictness consciously.

Classroom assessment, one of the most common formative assessment techniques is different from summative assessment in this respect. One of the chief aims of formative assessment is to provide guidance for the teacher by investigating the process of learning by providing continuous assessment. In other words, formative assessment is applied to adjust the teaching process and students' learning processes to the instructional goals with the aim of maximizing students' achievement. Raters appear to be aware of this distinction between formative and summative assessment, and seem to apply the different forms of evaluation adapted to suit their instructional purposes.

Raters usually assess their own level of strictness based on external, rather than internal signs. The notion of strictness is not a quality developed internally and consciously, but a characteristics prompted by external factors, a relative feature which can be assessed in comparison with other raters' strictness.

Rater 2

This turns out during the agreement procedure.

Rater 7

I am one of the strict markers. This usually turns out during the agreement procedure when it is me who always gives lower scores.

Rater 9

Well, my impression is that I am inclined to be stricter than the average. This is not what I would think of myself, not necessarily, but when we do the double marking it turns out…, it rarely happens that anyone is stricter than I am. They either give the same mark as I do, or give better mark, higher scores.

It is also interesting to compare how raters' feelings about their own strictness correspond to their actual practice. Strictness data from the FACETS analysis, the scores given on the three scripts marked by all raters and their own intuitive feelings will be compared next. Table 27 summarizes how raters' measured and perceived strictness are related to each other.

Table 27 A comparison of the fifteen raters' measured and perceived strictness

Rater	FACETS	Small sample	Interview
Rater1	NA	++	0
Rater2	+	-	0+
Rater3	-	+	0-
Rater4	-	--	0-
Rater5	+	-	+
Rater6	NA	-	0
Rater7	-	+	-
Rater8	-	--	-
Rater9	NA	0	-
Rater10	+	++	+
Rater11	NA	+	0
Rater12	+	+	0+
Rater13	-	+	0-
Rater14	NA	--	0+
Rater15	NA	+	-

+ = lenient
- = strict
0 = neutral, average
NA= no data available

The second column shows raters' strictness according to the FACETS data used in the current project. As it has been stated earlier, no FACETS data are available for each rater, which is clearly indicated in the table. + signs indicate leniency, - signs show harshness. In cases where rater characteristics can be considered extreme, these marks were doubled. The 0 sign indicates hesitation, in other words reluctance to admit openly and without lengthy explanations whether the rater considers himself or herself strict or lenient. The comparison of the data obtained from different sources shows an

interesting picture. Firstly, it should be noted that the FACETS and the interview data indicate a higher degree of correspondence in defining raters' leniency and harshness. There are noticeable differences between the strictness measures on the small sample and either the FACETS data or the interview data. This is not a surprising result, as a small sample size is more susceptible to individual variations and less sufficient to generalizable results. Out of the fifteen raters nine have a full profile with all three types of leniency and harshness data. Rater4, Rater8 are unanimously characterized as strict markers, and the different types of data also support the leniency of Rater10 and Rater12 in complete accord. In the case of the remaining six full-profile raters there is an agreement between the FACETS and the interview data in defining their leniency and harshness: whereas Rater2, Rater5 appear to be on the harsh side of the stringency continuum, conversely, Rater3, Rater7 and Rater13 seem to act leniently when marking writing papers. Raters' perceived strictness deserves further comments. More than half of the raters classified themselves as neutral which indicates their efforts rather than their actual practices. The data also show that after an initial hesitation marked by 0 in Table 27, raters were prepared to admit to their asserted strictness or leniency. There was only one exception, Rater8, who firmly, without the slightest hesitation defined herself as a strict rater. All in all, these results seem to suggest that raters are well aware of their level of strictness, which, in the majority of the cases, is confirmed by the numerical results. Nevertheless, they wish to see themselves as not belonging to either extreme.

As might have been expected, an investigation along the language division indicates that there is no complete overlap between raters' intuitions and their observed and measured strictness either in the English or in the German group. With the German raters (R4, R7, R13 and R15) and the German scripts there seems to be more

consistency in the data obtained from different sources. It appears that all four raters tend to be slightly stricter than the average, and they are also more or less aware of this.

> **Rater 7**
> I am one of the strict markers. This usually turns out during the agreement procedure when it is me who always gives lower scores.
>
> **Rater 4**
> I am somewhere in the middle. Neither strict, nor lenient. … But maybe I am a bit harsher than the others.
>
> **Rater 13**
> I am strict in an average way. It is true that I am by no means lenient, but I insist on adhering to the assessment scale and mark accordingly. But on the whole, strictly. So I consider myself strict rather than lenient.
>
> **Rater 15**
> I am strict I think. I am definitely not lenient.

The English markers show a less balanced pattern, and the amount of data available allows more tentative generalizations about their perceived and actual strictness. Mention should be made of the higher degree of similarity of supposed and measured strictness when the interview data are compared with the results obtained with the FACETS analysis. The FACETS data in Study 1 are obtained from a larger number of scripts marked and thus provide more reliable and generalizable data than those obtained from the marking of three scripts only in Study 2, even if those scripts were all marked by all raters. In addition, as markers remarked earlier, they need time and an adequate amount of papers to mark to accommodate themselves in the rating process. For all that, there are raters where all three types of data point into one

direction, and this is so with those who acknowledged belonging to one of the two extremes. Rater 12, one of the most lenient raters according to both the FACETS data and the small sample data, showed a full awareness of this characteristics of hers.

> **Rater 12**
> I am in the middle, sometimes I am strict, and sometimes I am lenient. But on the whole I am not very strict.

The rater pair is mentioned in the interviews in relation to three issues: adjustment to the pair's strictness, behaviour during the agreement procedure and favoured and disfavoured rater pairs. Adjustment to the rater pair's strictness is not a typical characteristics in assessing writing performance. This is partly so as markers may not know who their rater pair will be. Even if they have preliminary information about their pair, they use their knowledge of the pair's rater characteristics during the agreement procedure. In other words, in the agreement procedure they make an effort to reach a decision on the final score based on the familiarity with the rater pair's strictness rather than adjust their own strictness to that of the pair during the marking process. The situation is slightly different with subjectively scored oral performances. Although the assessment of oral performances is not central to the focus of the study, it should be pointed out that some raters exhibit different characteristics depending on whether they mark writing or oral performances.

> **Rater 3**
> I wouldn't want to divert, but I have to mention the oral exams because there I am unfortunately influenced by my pair. I know that this is highly undesirable, but there is some kind of automated adjustment in my thinking. If my rater pair is stricter than me than my stricter self will act and vice versa. And I wouldn't

say that this is because I want to avoid conflict, because I am ready to fight for a score I find fair.

The following comment was made by the English Chief Examiner who organizes the standardization and marking sessions and has a greater familiarity with the typical rating behaviour of all raters.

Rater 14

I have seen quite enough statistics to know who is strict and who is lenient. But even without that after so many years I would know that. So instead of reaching an agreement I am trying to adjust the scores accordingly.

The agreement procedure is very closely associated with the notion of "conflict" in the raters' minds. Nine participants used the word "conflict" in some form in the interviews, and it was used altogether fourteen times. The agreement procedure whereby the two raters seek to reach a consensus concerning the final score does indeed involve an element of conflict. Whereas the source of this conflict is apparently the difference between the awarded scores, it can eventually be put down to different personalities and attitudes.

Rater 2

I would definitely go into conflict for the sake of the candidate. I would never say it is all the same for me or resign easily.

Rater 10

There are two people I …I would rather not go into conflict with… so I let them do as they please.

Rater 14

In adopting strategies during the agreement procedure, I have to use different strategies with different people. There is for example XY *(gives the name of the rater*) when I have to be more assertive and defend my position more strenuously. And there is XY *(names a rater again*) for example, who tries to avoid conflict and leaves the decision to the rater pair. This again calls for a different strategy.

The issue of conflict also arises with differences between global and analytic scores. As it was also apparent in the think aloud protocols, some raters are often faced with the dilemma of having to diminish or do away with the discrepancy between the scores allocated according to the analytical assessment scale and their intuitively set global scores. This rater behaviour, however, is not common for all raters. Whereas some try to strike a balance between their intuitions and the score informed by the analytical scale, for others adding up the subscore is just a rating technicality, and the total score carries no special meaning. For one rater, this conflict was a problem at the beginning of her rater career but does not present problems after several years of marking experience.

Rater 4

I know a colleague who in such a situation would say, Oh, no, this can't be 17 points. But this is not my style. The final score is what is added up from the subscores and that's that.

Rater 6

When we started, at the very beginning, it did happen that I was quite surprised to see the final score. Which was usually more than what I would have given intuitively and globally. Globally we can give scores from 1 to 20, or rather from zero to 20, and then I found that according to the assessment scale I had to

give more points than I would have thought. Because of the assessment criteria. Maybe then, at that time, I had a second look at the script. By now I have got used to what is quite difficult to accept that a performance which would deserve a mark one (*according to the general 1-5 marking scale used in Hungary*) can get 10 points. But in fact 10 points is below the pass mark, so it means that the target level has not been reached. I don't think there is a conflict between the analytical and global score, not now.

Rater 10

After reading a script, I usually have a global score in mind. And my intuitive global score is better than tearing the performance into pieces according to the criteria.

Rater 13

I prefer analytical scales because I can concentrate on different aspects of the performance in a more focused way and the criteria do not influence each other, in my judgement, that is. Of course there is bound to be some overlap between the criteria, but if I had a five point global scale with everything crammed into it, that wouldn't result in such a clear judgement as with the analytical scales.

Contrary to the results of the analysis of the think aloud protocol data, the raters, with one exception, are less willing to admit to carrying out post-hoc score modifications or adjustments.

Rater 13

There are cases when I am slightly surprised at the total score. Then I go back to the criteria and check whether I have used them correctly. And if it turns out that I have used them correctly, then there's nothing doing. And if it is really nasty, I check where I can deduct points. Or make sure that a letter that deserves 10 points shouldn't be 12 points.

One comment deserves special attention. For one of the interviewees there is one global criterion, whether the letter produced could be put in an envelope and sent it to the addressee. The global criterion is life-likeness or authenticity. This rater transforms the simulated authenticity of the exam into real in her assessment practice. For her task achievement means usability in real life and thus removes the simulative nature of the examination.

> **Rater 2**
> Maybe I have one global criterion: could I put this letter into an envelope and post it? Post it together with its mistakes. So we can have a look at the mistakes individually and one by one but also globally. And because of this global criterion, we consider in this business genre whether this would work in real business life.

It is not customary in Hungary to base the assessment procedure of an examination system on pure, non-scored mastery and non-mastery decisions even if no legislative limitations would prohibit such a practice. This rater's attitude, however, suggests such an approach.

7.1.2 Extremism and central tendency

Extremism and central tendency are considered the types of rater misbehaviour whereby raters deviate from the fair score by either tending to show preference towards either end of the scale, or alternatively, by showing lack of variance in the scores and overusing the middle categories. The interview questions inquired about the frequency of occurrence of these categories in the raters' rating history as well as the possible underlying reasons for sticking to the safe middle category.

7.1.2.1 Maximum score

What seems to be an intriguing issue is whether a maximum score assumes an impeccable piece of writing without any mistakes. It seems from the raters' comments that there are a large number of requirements that are needed for the maximum score, and the complete lack of mistakes is not necessarily the most important one among those. For the majority of the raters there might be some minor mistakes in a performance with the maximum score, for a few, however, perfection is the norm. It is also apparent that raters' emotional attitudes to the notion of maximum score play an important role in assigning a maximum score to a writing performance. The following comments reveal rater attitudes according to which minor mistakes are acceptable as long as they are in conformity with the assessment scale.

Rater 1

A test performance with a maximum score is not necessarily without any mistakes altogether, but there are exactly as many mistakes as the descriptor band allows for. So if there are just very few mistakes or a couple of slips … of course because of the imperfectness it is difficult to decide, but I have no objections to giving the maximum score.

Rater 2

When I give maximum score and my rater pair doesn't, my main argument is that this is an intermediate exam. So what is the maximum at intermediate level? Is it the lower bound of advanced level? Is a maximum score at intermediate level a lower advanced level? No, I don't think that a maximum score intermediate would pass for a weak advanced level. A maximum score intermediate level is intermediate level. And because of this, it is not necessarily perfect. My yardstick is not the native level and I don't think that it is a good thing that we let ourselves be pushed into that direction. Because I think that there is such a thing as an absolute intermediate level. There might be mistakes in a written performance with 20 points.

There are raters on the other end of the scale who are less willing to tolerate mistakes and openly admit their high expectations. Out of the fifteen raters interviewed, two were absolutely resolute about the maximum score, and the third also tended to prefer perfect performances for the total score, but s/he was slightly more flexible. Interestingly, the two perfectionist raters who seem to have extreme expectations for the maximum score were both German raters.

> **Rater 13**
> A maximum score of 20. I don't think I have ever given 20 points. Because I am perfectionist and strict.

> **Rater 7**
> You can give maximum score for a perfect performance only.

> **Rater 5**
> I very seldom give maximum scores. It happens quite frequently that there are a couple of minor mistakes in the text and then I give 19 points. Mistakes which I myself would also probably make in an examination situation. Inappropriate use of words or something like that. And then my colleagues tell me off and say, look this is advanced level. And then I say that even so, there are three mistakes in it. But in the end they manage to convince me. But if I feel that it is a mistake and not just a slip, I will fight for 19 points instead of the maximum score.

Apart from the number of mistakes, their types are also commented on. As we have seen so far, errors and mistakes or slips are systematically differentiated. Furthermore, raters find it important to link the number of mistakes with text complexity. They seem to be more permissive if the candidate attempts to produce a more complex text and

thus is more likely to make mistakes, and conversely, they seem to demonstrate less preference towards error-free but simplistic pieces of writing.

> **Rater 2**
> If it is completely free of errors but the candidate uses primitive structures, it is not maximum score, but if, say, the structures are more complex and there are a few minor mistakes here and there, I would happily give maximum score for that. So it is not the perfect perfect that is perfect…

Unlike perfectionist raters who penalize even the tiniest slips, there are raters who tend to mark performances up for emotional reasons. As they do not feel much difference between 19 or 20 points, they are willing to opt for the maximum score for the emotional surplus and sense of success the maximum score gives to the candidate.

> **Rater 3**
> I don't think I have given maximum score too often but definitely more frequently than zero score. If I find that a letter is around 19 points then I have a very strong desire to give the maximum score. Probably I am inclined to give 20 points when the performance would have deserved 19 only, because as with zero and 1, there isn't much difference between 19 and 20, the only difference is that the total score will probably make the candidate happy. And why not make them happy?

> **Rater 14**
> The problem is that you always have the feeling that it could have been done even better. But it does happen that someone gets 19 points and you feel like giving 20, let's have a perfect one, and let's acknowledge that nothing better can be expected than this.

7.1.2.2 Zero score

Almost all raters agree that zero scores are highly infrequent. This is partly due to the fact that if there is one subtest with zero score, the complete test is a fail regardless of the rest of the results. A full zero, that is, zero on all four criteria is only conceivable if the task is missing completely or no single letter has been put on the answer sheet. One zero score on any of the criteria is more likely but is also very rare. As raters remark, a zero score might have undesired psychological implications which can be avoided by giving 1 point instead of zero. A very low score conveys the implied message, yet does not humiliate the candidate. This is even so if giving 1 point instead of zero means deviating from the scale. The human aspect of rating and its implication on the candidates are very important for the markers, possibly more important than sticking to the assessment scale and using the categories in an appropriate fashion.

> **Rater 3**
>
> Honestly, I don't think I have ever given zero point. Neither when the total is zero, but nor even when one subscore is zero. I insist on giving zero point only for a blank sheet. And this is not because I want to avoid conflict. What I think basically is that someone who deserves zero point will fail anyway, and there is not much difference between 1 or 2 points from the perspective of the final result. The impact, however, of a zero score or 1 point on the candidate will be quite different. A colleague of mine told me something very wise once, something that I always bear in mind while making decisions that the candidate can be failed but should not be humiliated. So I think I can fail the candidate with 1 point but at least I don't humiliate him/her.

Another rater goes as far as to hypothesize that a candidate who performs very badly on the writing task is highly likely to fail the rest of the test. Although s/he is generous enough to give 1 point on the writing performance that would deserve to be zero, the

achievement on the rest of the test will probably be not enough for the candidate to pass.

Rater 5

> The letter is extremely informative. If someone cannot put a proper letter together, it is highly likely that in the test s/he will get 5 or 6 items correct by chance out of the twenty-five. And in the oral exam s/he will definitely fail, so because the test consists of several parts I am inclined to give, say 3 points, because the candidate will fail anyway.

Another rater perspective is that even within the fail category differentiation between the extremely low scores carries information for the candidate and provides useful feedback on their performance.

Rater 2

> I carefully consider whether it should be 2 or 3 points because this provides information for the candidate. When next time, after working hard and preparing for the exam still gets lower score than on the first occasion then s/he will commit suicide. I always have my darling weak students in mind who indeed make an effort… so there is a human being behind every effort, and I know that it is important that if on the first occasion they got 4 points, next time they should get 7. It is feedback.

The zero category presents a serious and seemingly insoluble problem to raters. In cases when the task is misunderstood altogether, the assessment criteria do not make it possible to give a total of zero for that performance. Whereas on the Task achievement criterion the zero score seems fully justified, the other criteria are more problematic. The dilemma is whether an impeccable piece of writing which deserves full scores on the remaining criteria can be evaluated on those criteria when the aim that the task

requires has not been achieved. If such a piece of writing gets a total of zero, it can be rightly claimed that the task achievement has an overarching effect, and it contaminates the other criteria. The question then arises what the purpose of an analytical rating scale is in such a situation. On the other hand, if such a piece of writing is accepted as a pass, any memorized piece of text could be accepted for any type of task. Arguments could be made both for and against failing such a performance altogether, but this is not the point here. Whichever decision is made, namely whether only one criterion should be zero or all in such situations, all raters should act and rate accordingly.

Rater 4

For me there is no difference between a blank sheet and a piece of writing which is totally different from what is required by the task. The candidate received a task which is related to the expected knowledge area. So I wouldn't deal with a performance at all which is about something different. This is not the commonly followed practice, but for me it has always been a big question and an issue for debate. Can I learn Winnie-the-Pooh by heart, put it down as the writing performance regardless of what the task requires and I get a language certificate for it?

Rater 7

It is very difficult to give zero score according to the assessment scale. It might only be possible with the task achievement criterion if the candidate misunderstood the task and say, wrote an offer instead of a request for an offer. But it is extremely difficult to give zero on the other criteria, because there is vocabulary to assess even if it is a request for an offer but an offer and the vocabulary will be similar specialized vocabulary even if it is slightly different. There will be language use to assess, so we are pressed to give some points on that criterion and the situation is the same with style.

Rater 9

I don't think there is such a thing as a total zero. If the candidate writes something sensible about something different … then the s/he has produced something. Yes, but on the other hand, extremely weak performances deserve the same as blank answer sheets.

Raters are rather indecisive about zero scores and performances which are altogether different from the expected task. All the quoted excerpts confirm the need for a consensus between raters regarding the issue of task misunderstanding. Whatever decision is made all raters should stick to it, and the practice should be made uniform across similar tasks and across examination periods.

7.1.2.3 Central tendency

Attitudes towards the central categories and the underlying reasons show a heterogeneous picture. As might have been expected, for some raters the central categories serve as a safe middle of the road attitude, especially for reasons of fatigue. Apart from tiredness, hesitations leading to opting for central categories might be due to lack of enough samples to conveniently locate the script in its proper place in the rating scale. This is especially typical at the initial stage of marking.

Rater 1

I am fully aware that this is something to fight against. (*Using central categories*). In such cases the assessment scale should be read over and over again. It does happen indeed. There are cases when one gets lost and starts to be indecisive. This is either when you haven't marked enough scripts yet or when you have done far too many.

Rater 5

It depends on the task. When it is enough to lift the prompts and the candidate has to do nothing but slightly change what has been given, insert an auxiliary and put a memorized chunk at the beginning and at the end of the text, then I can't really do much. There's nothing to underline. And we have this ingrained feeling that when there is much red ink then the candidate gets few points, when there isn't, it means lots of points. So central categories come when there aren't many mistakes in the text, but not much creativity is shown either.

One rater remarked that the central categories in the four criteria result exactly in the pass level of 60%, so playing it safe and using the central categories does not actually disadvantage the candidate.

Rater 7

I have come to realize recently – although not necessarily in recent marking sessions – that if you are pressed for time, and that was especially typical when there were also entrance examination papers to be marked that after the twentieth letter you can't really think clearly enough. I think that after a certain number of letters there should be a break inserted in the schedule. It is in these cases that I check whether the script deserves to be at 60% or not.

Interestingly, for some raters the middle categories do not serve as a recourse, as they appear to be the most problematic to apply. It is difficult to decide what deserves to be an average performance, for them assessing extremes, outstandingly good our bad performances are more straightforward.

Rater 11

There is this "hesitation in the middle". When you have seen many of all kinds of scripts. Then you start hesitating in the middle. And it happens in the middle of the scale. Obviously it is much easier to use the top or the bottom end of the scale. You have to consider the middle part of the scale very carefully.

The above comments confirmed that raters adopt different attitudes to the extreme categories. Zero for the total score is almost a practical impossibility. The only clear and unquestionable instance when a total of zero is justified is the complete absence of the task. A quandary which needs further consideration is the misunderstanding of the task and deciding how such a performance should be assessed. It also seemed from raters' remarks that they do not appear to be at ease with extreme categories, the frequency of their application is below the average. Raters also mentioned the possible emotional implications of giving extreme scores. It should be emphasized that this project seeks to identify rater misbehaviour and what has been so far found in relation to the extreme categories is the justification of using them rather than instances of misusing them. As for the overuse of the central category, although two raters pointed out the difficulty of its interpretation, the majority of the raters linked this rater misbehaviour to rating process technicalities: marking either too few or too many papers might result in the overuse of the central categories.

7.1.3 Halo effect

Several instances of the halo effect, the cross-contamination of descriptor bands have already been discussed. The next section focuses more directly on how the rating criteria influence each other and which criteria are more important and less useful from the perspective of the assessment of writing performance. At the time of constructing and validating the interview protocol, it turned out that the halo effect might be the most difficult rater misbehaviour to identify. To find out about raters' attitudes to the assessment criteria, they were asked about the usefulness of the criteria, which criterion they found the most and the least important, and also about the ease of their application. At the final stage of the interview, they were also asked to rank order the criteria

according to importance and ease of application. This was meant to provoke an answer and push raters into providing an answer to the question and identify a hypothesized halo effect. It is a highly reassuring finding that raters were not easily manipulated in their answers, even when forced to take sides, they were reluctant to go against their beliefs. They did not feel any kind of hierarchy between the criteria, and they considered them equally important. This attempt failed as an element of the research method, but this very lack of the result confirmed the raters' proper attitude to the rating scale criteria.

> **Rater 4**
> After all, each criterion is important. All of them are important.

> **Rater 7**
> I don't think I would like to get rid of any of the criteria. They are all equally important.

Nevertheless, it appeared that one criterion, namely Task achievement, is possibly more all-embracing than the others.

> **Rater 7**
> Probably the task achievement criterion is the most important because it is about producing the appropriate type of writing. So I consider task achievement the most important but it cannot be viewed in isolation.

> **Rater 10**
> I consider task achievement very important. Because there is a task which should be completed. But it is a tricky issue because for me grammar and language use are also very important. So for me task achievement and language use are the most important.

A recurrent problem seems to be the misunderstanding of task. This came up in connection with the halo effect again. One rater specified what could be labelled as anti-halo effect, namely that the misunderstanding of the task which clearly belongs to the Task achievement criterion should not be penalized with regards to the other criteria.

> **Rater 15**
> Task achievement is quite a dominant criterion. I am trying to bring up a typical example. When the candidate wrote something different from what was expected, for example wrote an offer instead of a request for an offer and produced a stylistically perfect offer, which is a typical task at the intermediate level. And I have to give maximum score on style whereas s/he might get very low scores on the other criteria. But poor thing has learnt something, and we should appreciate that.

As for the connection between criteria, raters establish direct and indirect links between them. These connections do not necessarily imply a halo effect; they are only regarded as criteria more closely associated with each other than with others.

> **Rater 1**
> I think the point in using an analytical rating scale is that there shouldn't be any overlap between them. In spite of this …as we have seen before, it does happen that task achievement is not satisfactory, but language use is fine. If s/he gets too high a score for language use then the final score will contradict our global assessment. And, well, in such situations a kind of nasty score adjustment takes place. It can be perfect from the perspective of language use even if task achievement is problematic.

Rater 5

There is some inherent connection between them. Take language use and grammar, for example: fragmentary sentences are not appropriate to the genre and that usually goes together with a poor vocabulary. So for the two more subjective criteria *(task achievement, style)* the candidate will get higher scores, and for the two more objective ones *(vocabulary, language use)* where you can underline lots of things in red the candidate will get lower points.

An important comment made by some of the raters is that although the criteria are treated separately, it does not happen very often that a performance displays highly different characteristics on the different criteria. This remark is probably based on years of practice rather than on theoretical considerations.

Rater 4

Well, in actual practice if task achievement is 2 or 3 points, then it is difficult to imagine 5 points on the other criteria. I don't think I have ever seen such a thing.

Rater 9

It is not extremely typical that there are nasty grammatical mistakes in the text, and together with this, style is brilliant or the vocabulary is impressive.

Unlike the majority, in one rater's view the rating criteria are fully independent and there might be substantial differences between the subscores given on the individual criteria.

Rater 10

I don't think there is a connection between the criteria. The score can be easily 1 point on one criterion and 5 on the other. Easily.

For one rater there is connection between the criteria, but it stems from administrative rather than theoretical grounds. S/he treats two borderline scores together to make sure that with one criterion the candidate should get the higher score and with the other one the lower score.

> **Rater 14**
> I usually leave borderline cases, where I can't decide between two scores, to the end. If there are two criteria on which the scores seem to be borderline cases then I will make a decision considering both. I will make sure that the candidate does not get either the lower score or the higher on both criteria, but I will try to strike some kind of balance. I also use pluses and minuses which help during the agreement procedure to better defend my position.

Although the halo effect is termed as rater misbehaviour, but raters claim that there is a natural, inherent relationship between the criteria, and in certain cases it is very difficult if not impossible to view them separately.

> **Rater 12**
> I think task achievement is the most important criterion. And style. It is difficult to answer this question because they are all related to each other, the former two with vocabulary and language use. Because if the vocabulary is inadequate or very poor then style cannot be highly satisfactory, either.

There are very few instances of acknowledged halo effect when it can in fact be regarded as misbehaviour. One rater mentions the negative dominant role of task achievement which she considers a practice to fight against.

> **Rater 9**
> I don't consider it fair practice to mark up a piece of writing if the task achievement is exceedingly good. This is what I have experienced: good task

achievement has such an effect that it encourages the marker to overestimate all criteria. I don't think this is good.

Rater 6

Language use … I am surely influenced by language use. I cannot really spot it, but I can sense it. So probably if the text is horrible grammatically, whatever beauteous phrases and expressions are used, they wouldn't influence me. But I don't actually directly perceive that. I don't feel it that directly.

The existence of the halo effect is more evident in the case of oral examinations, according to one of the interviewees. S/he assumes that the assessors are quite likely to mark down a weak oral performance on all criteria without due consideration even if the candidate performs badly only on one criterion.

Rater 9

It is typical in the oral exam that when a candidate is weak then the examiners are inclined to give low scores on all criteria, for example low score on comprehension, although the candidate's comprehension is not worse than that of the others' but to make sure the final score is low they give low scores on all criteria.

Finally, one rater noted the type of halo effect when it is not a certain criterion but a certain score that seems to contaminate the other scores.

Rater 13

I don't find this with any of the criteria but rather with the middle score. If I give the middle category for task achievement and vocabulary then somehow automatically I tend to give the middle category for the other criteria as well.

From what appears in raters' views, the halo effect seems to exist in different forms. With one exception, all raters acknowledged the existence of some kind of connection between the rating criteria even if theoretically they should be treated completely separately. The most frequently mentioned connection is between the task achievement criterion and style. It should be pointed out that the majority of raters did not admit to experiencing a negative dominant influence of one criterion, as they only established an inherently existing link between those criteria. Another form of the halo effect harks back to an earlier issue, namely that complete misunderstanding of the task and a zero score for task achievement might justly contaminate the other criteria and nullify the candidates' chances to obtain scores on the remaining criteria thus nullifying the total score. The halo effect might also be related to a dominant score, and for some of the raters it is also more apparent in an oral examination.

7.1.4 Response sets and playing it safe

Response sets and the playing it safe rater strategy are not very common. Response sets are scores following some kind of regular pattern of numbers rather than scores related to the actual performance. In discussing potential halo effect, one marker claimed that especially the central category sometimes had a special attraction and encouraged him/her to give the same score on the other criteria. Another rater noted that regular patterns in the scores which are generated by chance caused him/her worries, as such a set of scores might suggest to the chief examiner potential carelessness or fading interest on the part of the rater. No other explicit appearance of this misbehaviour was detected, and neither did it emerge from implicit comments. Raters' playing it safe attitude would imply a form of behaviour whereby one marker awards scores close to the scores or identical with those given by the rater pair. Several objective factors make

it impossible for this rater behaviour to occur. Firstly, raters mark the scripts independently, and they often do not even know who the second rater will be. Secondly, one script is accompanied by three marking sheets: one for each individual rating by the two raters and one for the final agreed score. As these data are regularly checked and analysed, no rater would risk relying on the scores given by the other rater, or in other words copying them instead of carrying out proper marking individually. At this point it should be underlined that although conventionally raters are labelled as first and second rater, but practically they both act as first raters as regards scoring. As it has been said elsewhere, the actual marking of the mistakes is usually done by the rater who gets the script first, so it is only the amount of red markings that might influence the rater who is the second to evaluate the paper. This is not to say, however, that the playing it safe behaviour should not be dealt with as a theoretical possibility. During an earlier operational phase of the examination board such instances were disclosed and taken on with due attention. Anecdotal evidence suggests that the reason for such behaviour might be either lack of genuine interest in carrying out the marking task in the appropriate manner on the one hand, and indecisiveness on the other. This latter source of minimal effort for the independence of rating can be well counterbalanced by rater training and retraining.

7.1.5 Rating instability

The instability of rating refers to factors that influence rater behaviour in a random manner and result in unsystematic measurement error. Raters were asked about factors that might result in excessive strictness or undue generosity. In addition, they were questioned about their perceived rating consistency both within and across rating periods. Although the consequences of rating instability are similar to those of general

strictness and leniency, instability is considered to be a less permanent trait which appears irregularly and is more difficult to control for.

Next, potential sources of inconsistencies will be discussed based on the rater interviews. In the first study, measurement bias was explored as a psychometric quality; in the following section bias will be discussed in its psychological meaning. Both formal and content related factors might encourage raters to over- or underrate a writing performance. Outstanding creativity sometimes masks linguistically poor performances, and its ability to break the monotony of marking might lead raters to mark a performance up.

> **Rater 2**
> You are extremely tempted to appreciate creativity, highly inclined. You have to hold yourself back. Of course we can reward creativity through the assessment criteria. But if someone is inventive or funny, then you are liable to overrate it. It is badly written but at least I had a good laugh. So you have this urge to give slightly more points than deserved, but you have to be aware of this and set a limit.

> **Rater 5**
> Deviating from the scale depends on the task type. I am diverted upwards by creativity. Because there is usually a task and the item writer prepared a sample solution and the majority produce a writing performance similar to this sample. If there is someone who, staying within the given frame, has brilliant ideas without going too far, well, I quite like that.

Creativity as fantasy or imagination is a positive characteristics referred to by one of the raters.

Rater 15

In a writing task it might appear that the writer's imagination is clearly shown. But I think that should be evaluated. It belongs to the task achievement criterion.

Apart from creativity, intelligence which shines through the writing performance is also a characteristics which raters are likely to evaluate highly. Five raters out of the fifteen interviewed mentioned intelligence as a candidate characteristics likely to cause bias. Similarly, a pleasant personality which is manifested in the writing might also yield higher scores.

Rater 3

Clear, intelligent writing pampers my heart. Pleasant appearance also. Brightness is an additional plus. Such a positive global picture tends to mask a couple of weaknesses. I also value knowledge exceeding the expected level. As this is a further proof that the candidate is intelligent. But I don't think in this case it really matters that I mark the test up a bit against my will or conviction because such a candidate will pass easily anyway.

Rater 7

What is really heart-warming is an intelligently structured letter. If you feel that it is not just the replication of ready-made and memorized sentences, but the candidate does know about things.

Rater 13

If its layout is neat, well-structured. If you feel that s/he is intelligent, does not only repeat things, if you can sense creativity … and that s/he can do it.

Hypothesized positive human characteristics might also be a source of rater bias.

> **Rater 11**
> When you feel that you get something extra on the human side, when you feel that this is someone you would like to work with, then you might be a bit positively biased.

The most common issues raters seem to be positively biased to are related to human characteristics, such as sharp intelligence, great creativity and powerful imagination. Sources of bias related to formal characteristics include neat layout and clear handwriting. Whereas for one rater it is highly undesirable that there should be factors negatively biasing raters, others admit to having such inclinations. Interestingly, amongst the negatively biasing features, formal issues outnumber characteristics related to the candidate. One rater referred to a negative human feature, namely corruption, which s/he has very strong objections to and also attempts to fight against.

> **Rater 2**
> When a candidate writes about something dishonest, things like bribery or corruption. I know that my personal hobby horse is fighting corruption. I am inclined to move in a slightly more negative direction. Because I feel that corruption destroys everyone. If there wasn't corruption in the world, the general well-being of people as well as global economy could thrive everywhere.

Further admitted sources of biases are more to be associated with formal, linguistic features. Hand-writing and especially grammatical mistakes seem to be one of the chief sources of negative biases.

Rater 1

Illegible writing does upset me a bit. But not exceedingly. And anyway, handwriting is included in one of the assessment criteria.

Rater 5

Unclear handwriting drives me mad. Especially when there are lots of words crossed out in the texts. And also when it is not clear which is the draft and which is the final version. After I corrected half of the draft, I realize that it is not the final version. It is very difficult to remain objective in such situations.

Rater 9

I might be negatively influenced by unclear handwriting. Which might be considered in a way part of language use and grammar. Obviously such letters slow down the process of marking when you have to try and decipher the letters. But I am striving to be understanding, as I have children with illegible handwriting. So I am trying to convince my rater pair that this all will be done by word processors in the future, so there is not much point in being strict on that.

Grammar and language use are the most frequently mentioned areas where raters admittedly display a negatively biased attitude. What mitigates the impact of such hypothesized rater behaviour is that raters are aware of their differential treatment of the language use criterion.

Rater 5

When a candidate makes basic grammatical mistakes, I immediately start to feel that s/he doesn't speak English at all, only memorized a couple of sample sentences or used the dictionary and copied a few sentences from there, which, of course, she had previously carefully inserted between the lines in pencil. In such cases I can't help reading it like this. It is not about a few insignificant slips, but when I can see mistakes in basic grammatical structures. In such situations, in all likelihood, I tend underrate the performance.

Rater 6

The most frequent source of horror for me is almost always grammar, however much I feel ashamed of this. Although I don't know whether I should be ashamed of this, maybe it is more accurate to say that it is not fashionable to take grammatical mistakes seriously. It is always grammar that fills me with horror. Vocabulary cannot get that awful. A misused word might be funny or out of place. But vocabulary is less annoying, it is grammar that can upset me. If task achievement is not perfect, for example if the writer of the letter starts threatening or uses inadequate style it is a negative thing, but it does not actually upset me.

One interviewee goes as far as to specify the types of grammatical mistakes s/he finds hard to tolerate.

Rater 10

I think I am a bit grammar-oriented, which is not a very popular stance these days. There are things, such as an "s" ending , third person singular, or a past form. I know this is not fashionable today. A mistake in the third conditional doesn't make me very angry, but I don't think the "s" for the third person singular should be omitted. We use that ever so often, it should really be known.

Interviewees were also asked about their supposed rating consistency within one period and also across different periods. The general view concerning rater consistency was that it is easier to maintain within-period consistency, especially when few raters mark one type of task. In addition, raters pointed out again that leniency and harshness were largely relative characteristics and highly dependent on the actual task being marked. Thus, as tasks differ from period to period, it is more reasonable to talk about rating consistency within one period. With no exception, they felt that however strictly and carefully they attempted to follow the assessment scale, comparisons between

performances were inevitable and the rating context, namely the preceding performances indeed had an influence on the rating process. In answer to the question whether they would mark the same paper differently in two different rating priods, they felt that there might be only minor differences between the scores given on the two occasions.

Rater 1

However sad it might sound, there is bound to be some inconsistency, in the form of 1 or 2 points.

Rater 9

I am obviously influenced by the complete batch of scripts I mark. In a very weak pile a middling performance would get 13 or 14 instead of 12. But not more than that, because that would be a different category.

Rater 11

I am quite sure I would give different scores on the same script if it was put in a different pile. Perhaps 1 or 2 points more. Not more than that, though. By the way, the other day when we had to mark the same scripts for the second time, I gave almost exactly the same scores.

The importance of the standardization of marking emerged here again as a highly important element of the rating process. Also, markers' spot-checking and re-marking if necessary their own corrections to safeguard consistency is a common practice applied.

Rater 5

Standardization is extremely tiring and boring and also seems a total waste of time but I have only recently come to realize how important it is. When I had a second look at the first letter in my pile at the end of the day, I saw the 16 points

I gave and said, my God, for such performance at the end of the day I only gave 12 points. Then I went through these first few letters again, re-marked them and gave lower scores.

An additional form of rater misbehaviour, blackout, was mentioned almost by all raters. From a measurement perspective, it is also an inconsistency, but it can be clearly detected and accurately described. Blackout is an unusually high or low score awarded by one of the raters on a small set of performances, usually one, two, or three, where the differences between the two raters' scores are not consistent with the preceding differences. The existence of blackout provides additional evidence to the need of double marking which can diminish or eliminate its effect. The question arises what the consequences are if both raters experience blackout at the same time, with the same candidate. Although the likelihood of such a thing to occur is fairly low, such an occurrence might be the cause of a significant candidate/rater misfit in an IRT analysis. During the interviews one rater even preceded the question by referring to this special phenomenon experienced in the rating process.

Rater 2

I don't know whether you are going to ask me about this, but there is something which I experience, and I think this also could be a considered a kind of rater misbehaviour. It happens sometimes that in a batch of 100 papers there are always two or three papers where either on my part or my colleague's there is for example a five-point difference between our individual scores and then having another look at the script I shout out, oh my God, why have I given those points?

This rater also noted that the reason might have been that the rating process was disturbed or interrupted by some external circumstances. S/he also added that this

blackout did not necessarily result in extreme scores, rather in extreme differences between the two raters' points. Other raters mentioned fatigue as a possible source of blackout.

> **Rater 4**
> There is usually a period, which lasts for say five papers when there are extreme differences between the scores given by myself and my rater pair. One of us probably overslept or something. But I think this is natural because you do the marking from 9 in the morning till 6 in the evening practically with no break.

For another rater, the result of fatigue is that it over-sensitizes the rater to certain issues, such as an exceptionally bad start of the letter.

> **Rater 5**
> I would put this down to fatigue. I noticed this as early as the first rating session that there are rough patches. After reading the 40th letter there will be two, three or four when I can't understand in retrospect why I have given those scores. It is in these situations that subjective issues matter, such as awful handwriting, or something like "Dear Harry" or "Hi" to start the letter with. Unfortunately, I cannot take a business letter seriously when it starts with "Dear Harry" or "Hi" or something… And probably at this point I decide that this letter can't be more than 8 points and I probably won't even read it properly.

Another rater confirms that tiredness amplifies latent biases which are more difficult to control with fading interest due to great weariness.

> **Rater 6**
> This always happens and I attribute this to exhaustion that I pay attention in a different way. In such a situation we go over the scripts for a second time, and I don't quite see why we did it (*gave a certain score*). Or sometimes we might

actually find something that upset us, for example some horrendous grammatical mistake. But not always.

Such rater misbehaviour could typically result in unsystematic measurement error. As it has been pointed out earlier, the existence of a second marker eliminates the effect of this misbehaviour by bringing out unusual and unexpected differences, and encourages raters to go through the script for a second or third time if necessary.

7.2 Rater characteristics

Unlike other qualitative studies on rater behaviour (Vaughan, 1991; Lumley, 2005), the present inquiry attempted to retain the richness of data by avoiding the reduction of the phenomena to pure numbers. Whole lines of thoughts, chunks of texts were used to exemplify and illustrate certain typical instances of rating behaviour. The purpose of this type of analysis and presentation of the data was to give a deeper insight into rater behaviour which otherwise is highly problematic to capture. In addition, it turned out from the interviews that verbalizing seemingly automated activities and justifying them contribute to a more conscious approach to the rating activity and a heightened awareness of the possible consequences of rater misbehaviour.

According to raters, *rater leniency or harshness* is a relative feature, which may also depend on the task associated with the performance being marked. Raters usually did not consider themselves belonging to either extremes, and they were well aware of their level of strictness. They acknowledged the existence of factors which might increase or decrease their level of harshness, and which could be considered construct-irrelevant deviation. Among these factors, the ones which result in overdue leniency outnumber those which result in exceptional strictness. Creativity, intelligence, signs of

positive human characteristics which are filtered through the test task are features that might make raters overrate tests. Less important, but still significant score-distorting aspects include unclear handwriting. For several raters grammatical mistakes are a chief source of extraordinary strictness.

Extremism, although not a frequent form of rater misbehaviour, appears more often at the upper end of the rating scale when assessors mark performances up for affective considerations. For similar reasons, they try to avoid minimum scores.

Central tendency or overusing the middle categories was mentioned only once as a sign of tiredness. Applying *response sets*, a misbehaviour category into which central tendency could also fit, is a rater behaviour which is highly infrequent or almost non-existent amongst the raters observed and interviewed. Many of the raters claimed the analytical rating scale imposes stringent limits on the rating behaviour from which it was very difficult to deviate. Earlier findings, according to which analytical rating scales provide more guidance and allow a more focused concentration of several aspects of the performance, were also confirmed.

The *halo effect*, in other words the existence of one dominant assessment criterion appears to be a problem associated with one criterion, namely task achievement. Raters felt that this aspect was more overarching and encompassing than the other three, and thus was more strongly influenced by the others. Task achievement, although worded very carefully to differentiate its focal point from the other criteria, by its nature shows a marked interrelatedness with the other criteria. As for the relationship between other criteria, raters identified a naturally inherent connection between them, as they all belong to and describe one underlying language ability, and within that writing skill trait. Nevertheless, this intrinsic connection does not make it impossible to view and evaluate these features independently. The administration of the marking

procedure practically excludes the possibility of one rater consulting the rater pair's given marks, and thus the *playing it safe* rater strategy can be claimed to be non-existent. Perceived and admitted *inconsistencies* might be first and foremost attributed to exhaustion.

A special form of inconsistency was identified, which could be added to Linacre's list of rater misbehaviour. Blackout, observed by almost all raters, is a form of misbehaviour whereby raters show a highly inconsistent rating pattern for a small number of papers. Two possible causes were specified: one being extreme tiredness, and the other cause was an overwhelmingly negative impression at the very beginning of the writing performance. Such a feeling leads to an obvious bias, which prevents the marker from objectively evaluating the rest of the performance.

In the analysis of rater behaviour, some minor discrepancies have been disclosed. Apart from the confirmation of formerly identified forms of rater misbehaviour, the purpose of the study was to shed light on further aspects of inconsistent rating that might lead to measurement error. Based on the results, one important assumption should be made: what is termed as misbehaviour and conveys a negative connotation does not necessarily threaten the validity and the reliability of the rating process. The effects of such behaviour in all likelihood are less powerful and far-reaching in cases where markers involved are well aware of these problems. The interviews provide evidence for the awareness raters show of their own behaviour and rating characteristics, and also testify to their claimed efforts at eliminating these shortcomings. In addition, many of the potential negative consequences of rater misbehaviour can be cancelled out by double marking. Furthermore, the crucial role of rater training and retraining together with the standardization of marking received further confirmation. Whereas direct involvement and individual feedback on rater

performance might not necessarily fulfil its expected positive potential, a tentative familiarization with possible rater misbehaviour might definitely have a beneficial effect on raters' assessment practice.

Chapter 8: Conclusion

8.1 Validity of the rating

The aim of my study investigating rater and rating scale interaction and within that rater behaviour was to explore existing forms of rater misbehaviour and identify their possible sources. In the validation of the rating process apart from validating the rating scale, it is also essential to examine how the rating scale is used, and how valid the rating process is. It might seem an unorthodox idea to talk about rater validity, but how markers use the assessment scale and the extent to which they adhere to it or deviate from it might be rightly termed rater validity. With a full awareness of the basic difference between reliability and validity, rater validity is conceived as part of scoring validity (Weir, 2005) which is an overarching term for test reliability, internal consistency, marker reliability, and which

> concerns the extent to which test results are stable over time, consistent in terms of content sampling and free from bias. In other words, it accounts for the degree to which examination marks are free from errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate (Weir, 2005, p.23).

The purpose of my mixed-method inquiry was to provide empirical data to confirm the validity of the assessment of the intermediate writing task constituting part of the exam suite of the Foreign Language Examination Centre of the Budapest Business School. The direct target of the observation focused on two components of the rating process, the rating scale and the rater. The former was investigated by means of quantitative methods, whereas a broader understanding of the latter was expected to be

obtained with qualitative tools. The aim of the validation process was not to focus of the construct of writing, but rather on the validity of the rating scale use from a psychometric perspective.

The first broad research question sought evidence for the proper functioning of the rating scale.

1. Which assessment criteria generate bias of rater behaviour?

The FACETS analysis of the scores awarded on a six-point analytic assessment scale across an extended period of three years and including two languages yielded results identifying bias terms, but no consistent pattern could be detected in the data which would support the existence of systematic bias towards any of the rating criteria. This is very much consistent with earlier findings discussing rater related biases (Kondo-Brown, 2002; O'Sullivan & Rignall, 2001). The type of analysis carried out identified all types of biases, even insignificant ones. Although no consistent bias, that is, systematic error could be detected on the part of any of the raters, this kind of analysis should be regularly carried out as even insignificant biases might be informative. When discussing rater bias, it should be noted that bias, the average difference between observed and expected score might be either negative or positive, meaning that the rater is either too harsh or too lenient on the given criterion. When interpreting bias results, of course significant biases should be dealt with regardless of whether they advantage or disadvantage the candidate. At the same time a tentative suggestion is made that those biases which disadvantage candidates should be attended to with considerable concern. In the current project bias analysis extended only to the rating criteria, but to obtain a more comprehensive picture of the raters, it seems

reasonable to introduce this kind of investigation in other areas of subjective assessment as well. Rater and task interaction, and more importantly the results of task and specialization interaction could usefully be fed into the test development process.

The second research question examined whether raters use all criteria to differentiate between various aspects of writing performance.

2. Which criteria elicit little variation in the distribution of the awarded scores?

Although the formulation of the research question itself hypothesized a small range of scores associated with a certain criterion, the results were similar to those obtained in relation to biases. No regular inconsistencies or permanent category effect was apparent. The rating scale criteria were analysed with the help of FACETS, and the criteria measurement report provided data about category fit. In the datasets analysed, all data were within the acceptable range of infit, and no significant lack of variation or excess variation could be detected related to any of the rating criteria. An interesting finding is that there is no consistency in raters' attitudes to the categories in terms of leniency and strictness. There is no one single category which was consistently more difficult to get higher points on than on the others. A similar finding is reported by Eckes (2005), who found that although raters were consistent in their overall strictness, their severity appeared to be less consistent in relation to the rating criteria. This suggests that raters' interpretation and the associated strictness and leniency is probably highly dependent on the task and confirms raters' individual and personal understanding of the rating scale, which also appears to be situation-dependent.

The third and fourth research questions were both related to the proper functioning of the rating scale categories and the scale steps.

3. To what extent is the halo effect or the cross-contamination of descriptor bands apparent in the distribution of scores?

The halo effect is apparent when, in spite of using an analytical rating scale, markers rate holistically rather than analytically separating the different criteria. Low infit values suggest muted rating patterns when the same scores are given across all criteria. For all the cases examined, infit values ranged between .66 and 1.53. These two values were the only ones outside the acceptable boundaries, no other value showed muted or noisy patterns. The high reliabilities of the separation indices also confirmed that raters are capable of differentiating between the rating categories. Although among the results one relatively low infit value could be detected (.66) in the case of one rater, this does not indicate a general tendency towards the halo effect, and confirms that the different criteria are adequately applied by the raters. When discussing halo effect, two types should be differentiated, true halo effect and illusory halo (Engelhard, 1994). Whereas the former is a just and deserved score which shows a regular pattern, an illusory halo reflects in fact rater misbehaviour, whereby the rater evaluates holistically rather than analytically. Regular low infit values on the part of the same rater would confirm illusory halo caused by undesired rater misbehaviour, which was not the case in the present observation.

4. Does the factor structure of the total scores confirm the appropriate functioning of the 6-point analytic rating scale?

A FACETS analysis confirmed the appropriate functioning of the six steps of the rating scale. The scale is of the generally applied type, when higher abilities are associated with higher scores, and a weaker performance is linked to a lower score. Consequently, in the analysis the lower categories were expected to yield low logit values and vice versa. Such a statement might appear an obvious and unquestionable truth, yet this remains only a hypothesized assumption before it is empirically confirmed. The empirical data confirmed the appropriate operation of the six-point rating scale. Both the numerical and the graphic data testify to the existence of the six well-identifiable categories, in other words, the steps of the scale, which constitute the scores from zero to 5. Lower logit values were in fact associated with lower category scores, and higher logit values characterized higher category scores. There also seemed to be a gradual progression between the scale steps, and with one exception in the complete dataset always reaching the expected 1.4 difference between two categories. The reliability figures of the separation indices also confirm the separability of the scale steps.

Whereas the first group of research questions examined the validity of the rating scale from a psychometric perspective, the second major research question sought to explore rater and rating scale interaction, and identify sources of unusual rater behaviour. The observation of rater behaviour during the rating process with the help of data obtained from concurrent verbal protocols together with the analysis of perceived rater behaviour with interviews promoted a better understanding of rater practices and was expected to reveal possible sources undesired variability.

Why do assessors exhibit different rating profiles across different domains of the rating scale?

1. The numerical data in Study 1 did not confirm differential criterion functioning, or in other words, that raters attribute unequal attention to the criteria. The interviews, however, suggested that according to raters' perceptions, two criteria deserve special attention. The Task achievement acts as an overarching criterion which is difficult to view in isolation from the others. Equally important is to handle the Language use criterion with special care because for some raters admittedly this criterion may exert an undesired negative effect on the other criteria and the assessment of the performance. On the other hand, the fact that raters felt all criteria to be equally important and were unwilling to rank order them according to their significance indicates that they are aware of the equal importance of all criteria, and the numerical data testify that they act accordingly. This would at least partly refute McNamara's (1990) and Lumley (2005) claim that grammar is the dominant criterion in the assessment of writing performances.

2. What construct-irrelevant factors emerge during the application of the rating scale?

In Chapter 7 possible sources of rater misbehaviour have been listed. Differences in severity and leniency between raters exist, which is one of the fundamental rater characteristics, often claimed but less frequently confirmed empirically. This result is similar to the finding of Lee and Kantor (2003), who likewise concluded that raters are not equally severe, and thus they are not interchangeable. Although no unexceptional extremes were apparent, even minor differences in rater generosity and harshness are important to note. Although IRT approaches would make it possible to adjust scores for differences in rater harshness, this is a delicate issue which requires careful consideration. At the moment, instead of adjusting raw scores for

rater leniency and harshness and making the scoring system less transparent for candidates, the information on rater characteristics and behaviour is fed into the test administration procedures. If one rater shows a general trend towards either leniency or harshness, he or she should be paired with an assessor of a different rating profile. It should be avoided that either two raters with the highest or the lowest logit values constitute one pair. The regular rotation of raters might ensure an even spread of the differences in rater leniency and harshness and thus can be guaranteed that no candidate is disadvantaged by being assessed by an unusually lenient or harsh rater pair. Detailed information about rater characteristics might contribute to the creation of a rater bank, as Engelhard (1992) also suggested. Lunz, Wright and Linacre (1990) came to the same conclusion by emphasizing the need for "calibrated pools of items and judges" (p.13). Special care is taken that to maintain a balance to safeguard that no two raters with extreme rating qualities should be paired either in the marking process or in oral examinations. This approach might seem less professional than score adjustment, but it seems an acceptable compromise between psychometric perfection and the transparency of the scoring system made accessible to candidates. Other types of rater misbehaviour can be largely counterbalanced by double marking and the standardization of marking preceding the actual marking sessions. This urgent need for reorientation prior to each marking session is also pointed out by Lumley (2005).

Inconsistencies, which also lead to measurement error, have also been identified. It seems from the interviews that there are more factors which are conducive to a generous rater attitude than factors that trigger a negative approach. Although no deviation of the norm should be regarded as acceptable, it is tentatively suggested that a tendency towards more positive rater behaviour is less detrimental to the rating process. Pearson and Nayman's (1930) original idea of type I and type II error in hypothesis

testing, and the idea of false positives and false negatives are frequently applied in language testing. A false positive would be a non-master who, due to measurement error, is considered a master, whereas a false negative is a master as a result of measurement error rather than mastery. Raters' generosity might increase the number of false positives, whereas harshness would contribute to the emergence of false negatives. Neither of them are desirable in a valid and reliable testing context, but the existence of false positives can be considered in a way less unfair than that of false negatives. In other words, unduly rewarding candidates, even if not intentionally, is less harmful ethically than unjustly disadvantaging them. On the positive side, raters are susceptible to displaying an unduly generous attitude to signs of intelligence, positive human characteristics as well as creativity. On the negative side, markers tend to show oversensitivity to candidates' use of memorized chunks of language and prefabricated formulae. These are all factors which might positively or negatively influence the rater behaviour.

The results generated two further questions which allow us to consider the implications of the research in practical terms.

3. In which aspect(s) of the rating scale should amendments be made?

All in all, the results of both studies seem to suggest that the six-point analytical rating scale used in the assessment of intermediate writing tasks is adequately functioning: the four criteria are clearly separable, and the six scale steps can be applied to make fair judgements on the writing performances. As for the criteria, both studies imply that the task achievement criterion is the only one which needs further investigation. Most biases, however infrequent, both according to the quantitative and

the qualitative data are related to this criterion. Although major amendments do not seem necessary, minor changes in the wording of some of the scale descriptors were suggested. These comments are related to relative modifiers, a fair comment which is line with claims that scale band descriptors should free-standing, and not dependent on previous of subsequent steps of the scale (Hawkey & Barker, 2004; North, 2000; Shaw, 2004a). Also, the lower end of the scales should be more explicit and more clearly define the difference between a zero score and 1 point.

4. What modifications in the assessment procedure would contribute to a more extensively shared understanding and interpretation of the assessment criteria?

As the rating scale cannot be viewed in isolation and its usefulness and accuracy are the function of rater behaviour, the validation of the rating process should involve both the rater and the rating scale. The results seem to suggest that rater training tailored to individual rater characteristics and standardization may largely enhance the validity and the reliability of the rating process. The finding also seem to confirm that rater variability largely depends on the actual task being assessed, thus invalidating the general notion of a "trained rater". Initial rater training should concentrate on administrative and theoretical issues related to the marking process besides familiarization with the assessment scale and a simulated assessment practice. It is essential that each rating session should include a retraining session for the creation of the common frame of reference for marking the particular task in issue, making decisions regarding the extent to which task requirements should match the assessment criteria. In other words, consensus should be reached concerning what is expected and what is acceptable at a certain level. Information about rater characteristics should also be fed into the marking procedure. Raters of different levels of leniency and harshness

should be paired to exclude the possibility of creating highly reliable but also reliably extreme pairs. Raters showing inconsistencies, depending on their level of misfit, should either be omitted from the marking procedure or directed to the marking of objectively scored tasks. To eliminate the effect of "blackout" during rating, special attention should be paid to regular breaks that raters should insert in the marking process to ensure that no fatigue can contaminate the accuracy of rating.

The practical yields of the study indicate an urgent need to carry out a similar validation project for the subjectively scored oral performance tests. Whereas some of the findings are applicable to the assessment of speaking, a more in-depth inquiry is needed to investigate rater behaviour in the oral proficiency interview subtest. The results suggest that even though the standardization of the assessment of oral performances is an extremely demanding task and in resource-poor circumstances problematic to implement, to ensure fair and accurate rating of the speaking performances it is highly desirable.

8.2 Conclusion

The aim of the study was to validate a rating process including an operational rating scale and those raters who operate it in conformity with the guidelines set forth in the Standards for educational and psychological testing (1999), according to which "when previous research indicates that irrelevant variance could confound the domain definition underlying the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer "(p.46). Validity was

conceived as it is interpreted in modern test theory, "in terms of the statistical fit of each item to the model in a way that is independent of the sample distribution" (Wright & Masters, 1982, p.114). The validation took a strongly psychometric perspective: the rating process was investigated with Many-faceted Rasch analysis, and the results were submitted to further, qualitative analysis. The research questions emerged from these two areas: Study 1 investigated the quality of rating data over a three-year period, whereas Study 2 explored perceived rater behaviour. The development of a rating scale is a delicate process, and substantial amount of data are needed in order to validate what has been initially developed intuitively, by qualitative means or with the help of simulated data. As subjective, rater-mediated scoring is more susceptible to effects caused by construct irrelevant factors than objective assessment, only an ongoing in-depth monitoring of its use can ensure that it elicits valid decision-making. The validation of the rating scale is of prime importance in decreasing the subjectivity of performance assessment.

The study set out to investigate sources of measurement error with the aim of enhancing measurement precision and lessening the hypothesized inaccuracy associated with subjective assessment. Modern test theory, which makes it possible to decompose measurement error into random and systematic error, has informed the methodology of the research described. Instead of enumerating fundamental theoretical issues in language assessment at length, the review of empirical studies concentrated on features resulting in rater variability which might also contribute to measurement error. McNamara's (1996) theoretical model of performance assessment together with Engelhard's (1992) perspective of writing assessment served as the foundation of the theoretical background. Having consulted and summarized the most salient empirical findings related to the assessment of writing performance, certain focal points were

selected from previous studies to be fed into the design of the research. The analytical tool, Many-faceted Rasch measurement (Linacre, 1989) was also a basic cornerstone that shaped the research.

The results, which are based on data collected in a systematic way over an extended period of time, corroborate the validity of the rating process and provide empirical evidence on the adequate functioning of the rating scale and the raters operating it. The research failed to provide conclusive evidence on the existence of rater bias towards any of the rating criteria, which is a finding that attests to the validity of the rating process. The methods applied yield practical results in two major areas. Firstly, although no systematic bias could be detected in raters' use of the rating scale, the psychometric approach revealed minor, yet important deficiencies, and problems with the assessment scales that might require amendments. The research also confirms the need for the Multi-faceted Rasch analysis to be integrated into the test development process and become part of the ongoing validation procedure. Secondly, the identification of the sources of those deficiencies with the help of verbal protocol analysis and interviews provided invaluable insight into the nature of the rating process. Besides informing the rater training process as well as the standardization procedure, the results of such investigations help create a rater profile on which decisions concerning rater pairing should be based. The data obtained from these two sources might offer straightforward suggestions for enhancing rater efficiency and accuracy.

In sum, the results of the study provide convincing evidence that the existing validation methods should be complemented by those used and described in my research. Additionally, these procedures might serve as a potentially appropriate methodology and useful model for the validation of the more problematic and in many ways more intriguing testing of speaking skills.

Certain shortcomings of the research are nevertheless apparent: as the quantitative analysis required a special connected design and a certain sample size, it was not altogether possible to include all participants of Study 1 in Study 2 as well, and thus the first study identified certain forms of rater misbehaviour which were further explored with the help of other raters. This was meant to be compensated by collecting quantitative data from those taking part in Study 2, which actually did happen, but those data did not lend themselves to generalization about any kind of rater characteristics owing to the small sample size. Furthermore, lack of adequate sample sizes for the analysis also made it unfeasible to carry out the analysis including all languages in which examinations are administered. Finally, the project covered only the intermediate level which is the type of exam most frequently taken by candidates.

Concerning the original contribution of the present study to the field of language testing, it should be strongly maintained that the method applied for data analysis is definitely not new. The validation of the rating process, however, combining methods of modern test theory and qualitative means is not very common. An implied aim of the study lies in an attempt to promote Many-faceted Rasch measurement, this rather sparingly applied research method in educational research. This is an area where the present study hopes to add something new to the body of existing research.

As a conclusion, it seems timely to refer to Edgeworth (1890) again. We have indisputably come a long way since he made his claims about the problems related to educational measurement, in which he made a strong case for the importance of the investigation of the element of chance in assessment. His words, however, should always be borne in mind as a warning, and as a reminder of the need for continuous validation, as “however refined our methods of correction, we must not expect to

eliminate altogether the element of chance. There will remain an incorrigible minimum of uncertainty" (p. 663).

References

Alderson, J.C. (1991). Dis-sporting life. Response to Alastair Pollitt's paper: 'Giving students a sporting chance: Assessment by counting and judging.'. In J.C. Alderson, & B. North (Eds.), *Language testing in the 1990s: the communicative legacy* (pp. 60-70). London: Macmillan.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment. *Language Teaching, 35,* 79-113.

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 201-238.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: APA.

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

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer, & H. I. Brown (Eds.), *Test validity* (pp. 19-32). Washington, D.C.:Lawrence Erlbaum Associates.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

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

Bachman, L. F., & Eignor, D.R. (1997). Recent advances in quantitative test analysis. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language in education: Vol. 7. Language testing and assessment* (pp. 227-242). Dordrecht: Kluwer Academic Publishers.

Bachman, L. F., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12 (2),* 238-257.

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language in education: Vol. 7. Language testing and assessment* (pp. 275-287). Dordrecht: Kluwer Academic Publishers.

Barker, F., & Hawkey, R. (2004). Developing a common scale for the assessment of writing. *Assessing writing, 9(2),* 122-159.

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

Benke, E. (2004). Is it possible to measure teacher strictness? Traditional and modern statistical approaches to the rater facet in writing assessment. *NovELTy, 11(1)*, 36-45.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brown, A., & Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. In Huhta, A., V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96.* (pp.137-150). Jyväskylä: University of Jyväskylä.

Brown, J.D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced Language Testing*, Cambridge: Cambridge University Press

Buck, G., Kostin, I., & Morgan, R. (2002). *Examining the relationship of content to gender-based performance differences in advanced placement exams.* (College Board Research Report No. 2002-12.) New York: College Entrance Examination Board.

Budapesti Gazdasági Főiskola (2000). *Útmutató tesztfejlesztők számára.* (Unpublished internal document.)

Carmines, E. G., & Zeller, R. A. (1979). *Validity and reliability.* London: Sage Publications Ltd.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2(2)*, 155-163.

Choppin, B. H. (1989). *Item banking and the monitoring of achievement*. Slough: National Foundation for Educational Research.

Cohen, A. (1984). On Taking Language Tests: What the students report. *Language Testing, 1 (1),* 70-81

Congdon, P. J., & Mc Queen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37(2),* 163-178.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly*, *29(4),* 762-764.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Hott, Rinehart, and Winston.

Cronbach, L.J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507.). Washington, D.C.: American Council on Education.

Cronbach, L.J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Csapó, B. (2004). Tudásszintmérő tesztek. In I. Falus (Ed.), *Bevezetés a pedagógiai kutatás módszereibe* (pp. 277-316.). Budapest: Műszaki Könyvkiadó.

Cumming, A. (1990). Expertise in evaluating second-language compositions. *Language Testing, 7,* 31-51.

Cumming, A. (2002). Assessing L2 writing: alternative constructs and ethical dilemmas. *Assessing Writing, 8*, 73-83.

Cumming, A., & Berwick, R. (Eds.). (1995). *Validation in Language Testing*, Clevendon, England: Multilingual Matters.

Cumming, A. Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework.* (TOEFL Monograph Series Report No. 22). Princeton, NJ: Educational Testing Service.

Darwin, C. (1845). *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N.* 2d edition. London: John Murray. Retrieved 12 September, 2006 from

http://darwinonline.org.uk/content/frameset?itemID=F14&viewtype=text&pageseq=513&keywords=our%20poor%20of%20misery

Dávid, G. (2000). *The use of multitrak items and a small group oral in the context of the Hungarian education*. Unpublished doctoral dissertation, University of Glamorgan. Glamorgan, UK.

Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing, 24(1),* 65-98.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155-164). Westport, Ct: Ablex Publishing.

DeGruijter, D.N.M. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8(2),* 213–218.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly, 2(3)*, 197-221.

Edgeworth, F. Y. (1890). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society, 53*, 644-663.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian and Modern Greek. *Language Learning, 46(2),* 233-282.

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

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3), 171–191.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31(2),* 93–112.

Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* . (College Board Research Rep. No. 2003–1). New York: College Entrance Examination Board.

Erdosy, U.M. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. Retrieved August 6, 2004, from http://ftp.ets.org/pub/toefl/995738.pdf

Ericsson, K.A., & Simon, H. A. (1984). *Protocol analysis*. Cambridge, Massachusetts: The MIT Press.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Educational Limited.

Fulcher, G., & Davidson, F. (2007) *Language testing and assessment*. Oxon: Routledge.

A Gazdasági Szaknyelvi- és Módszertani Vizsgaközpont akkreditációs anyaga. (1999) Budapest: KVIFK.

Glaser, B.G. (1998). *Doing grounded theory: Issues and discussions.*: Mill Valley, CA: Sociology Press.

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

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw Hill.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing, *Assessing Writing,* 9(2), 122-159.

Henning, G. (1987). *A guide to language testing*. Boston: Heinle & Heinle Publishers.

Horváth, Gy. (1985). Tesztelmélet. Problémák, perspektívák. *Pszichológia, 1,* 53-78.

Horváth, Gy. (1991). *Az értelem mérése.* Budapest: Tankönyvkiadó.

Horváth, Gy. (1993). *Bevezetés a tesztelméletbe*. Budapest: Keraban Kiadó.

Horváth, Gy. (1997). *A modern tesztmodellek alkalmazása*. Budapest: Akadémia Kiadó.

Horváth, Gy. (2004). *A kérdőíves módszer*. Budapest: Műszaki Könykiadó.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18(1),* 89-114.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, *19*, 1-29.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741-746.

Kunnan, A.J. (1995). *Test taker characteristics and test performance: A structural modeling approach.* Cambridge: Cambridge University Press.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge, UK: Cambridge University Press.

Lazaraton, A. (2002). *A Qualitative Approach to the validation of Oral Language Tests.* Cambridge: Cambridge University Press.

Lee, Y., Breland, H., & Muraki, E. (2005). Comparability of TOEFL CBT Writing Prompts for Different Native Language Groups. *International Journal of Testing*, 5(2), 131-158.

Lee, Y., & Kantor, R. (2003). Investigating differential rater functioning for academic writing samples: an MFRM approach. Retrieved November 3, 2004, from http://www.ets.org/research/dload/ncme03-lee.pdf

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J.M. (1997). Guidelines for rating scales. MESA Research Note #2. Chicago, IL: MESA Press.

Linacre, J. M. (2003-6). *A user's guide to FACETS: Rasch-model computer programs.* Chicago: MESA Press

Linacre, J. M. (2006). FACETS. (Version 3.61.1) [Computer software]. Chicago:Winsteps.

Lumley, T. (2005). *Assessing second language writing*. Frankfurt am Main: Peter Lang.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12(1)*, 54-71.

Lunz, M.E., Wright, B.D., & Linacre, M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3(4),* 331-345.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Lynch, K. B. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University press.

MAXqda (Version 2) [Computer software]. Marburg: Verbi Gmbh.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7(1),* 52-75.

McNamara, T. (1996). *Measuring second language performance*. London: Longman

McNamara, T. (1997). Performance testing. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language in education: Vol. 7. Language testing and assessment* (pp. 131-139). Dordrecht: Kluwer Academic Publishers.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement, In H. Wainer, & H.I. Braun (Eds.), *Test validity* (pp. 33-46). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103*)*. New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Milanovic, M. & Saville, N. (1994). *An investigation of marking strategies using verbal protocols.* Paper presented at the 1994 Language Testing Research Colloquium, Washington, D.C.

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: theoretical concerns and analyses. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing* (pp. 15-38). Clevendon, Avon: Multilingual Matters.

Milanovic, M., Saville, N., & Shuhong, S.(1996). A study of the decision-making behaviour of composition markers. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 92-114). Cambridge: Cambridge University Press.

Miles, M.B., & Huberman, M.A. (1994). *Qualitative data analysis* (2nd ed.). London: Sage Publications Inc.

Molnár, Gy. (2003). Az ismeretek alkalmazásának vizsgálata modern tesztelméleti eszközökkel. *Magyar Pedagógia, 103(4),* 423-446.

Molnár, Gy. (2004). Hátrányos helyzetű diákok problémamegoldó gondolkodásának fejlettsége. *Magyar Pedagógia, 104(3),* 319-338.

Molnár, Gy. (2005). Az objektív mérés megvalósításának lehetősége: a Rasch-modell. *Iskolakultúra, 3,* 71-80.

Molnár, Gy. (2006). A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra, 12,* 99-113.

Myford, C. M., Marr, D., & Linacre, M. (1996). *Reader Calibration and Its Potential Role in Equating for the Test of Written English*. Retrieved August 28, 2004, from http://www.ets.org/ell/research/toeflresearch.html#rr52.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259

North, B. (2000). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*. New York: Peter Lang.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15(2),* 217-63.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Neill, T.R., & Lunz, M.E. (2000). A method to study rater severity across several administrations. In M. Wilson, & G. Engelhard, Jr. (Eds.), *Objective Measurement: Theory into Practice* (Vol. 5, pp. 135-146). Stamford, CT: Ablex.

Oppenheim, A. N. (2004). *Questionnaire design, interviewing and attitude measurement.* (Rev.ed.). London: Continuum.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*, 143-154.

O'Sullivan, B. (2006). Practical issues in language assessment. *Language Assessment Quarterly, 3(1),* 87-90.

O'Sullivan, B. & Rignall, M. (2001). *Assessing the value of multi-faceted Rasch bias analysis based feedback to raters for the IELTS Writing module.* Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report .

O'Sullivan, B. & Rignall, M. (2002). *A longitudinal analysis of the effect of feedback on rater performance on the IELTS General Training writing module*. Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report.

Pearson, E.S., & Neyman, J. (1967). On the Problem of Two Samples. In J. Neyman, & E. S. Pearson, *Joint Statistical Papers* (pp. 99-115). Cambridge: Cambridge University Press. (Original work published in 1930)

Pollitt, A. (1997). Rasch measurement in latent trait models. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language in education: Vol. 7. Language testing and assessment* (pp. 243-253). Dordrecht: Kluwer Academic Publishers.

Purpura, J. (1996). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Ryan, K., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9(1),* 12-29.

Saal, F. E., Downey, R. G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychometrical Bulletin, 88(2),* 413-428.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8(2),* 95-111.

Shaw, S. D (2002) IELTS writing: revising assessment criteria and scales. Phase 1. Retrieved August 6, 2004, from http://www.cambridge-efl.org/rs_notes/rs_nts9.pdf

Shaw, S. D. (2004a) IELTS writing: revising assessment criteria and scales. Phase 2. Retrieved August 6, 2004, from http://www.cambridge-efl.org/rs_notes/rs_nts15.pdf

Shaw, S. D. (2004b) IELTS writing: revising assessment criteria and scales. Phase 3. Retrieved August 6, 2004, from http://www.cambridge-efl.org/rs_notes/rs_nts16.pdf

Silverman, D. (2001). *Interpreting qualitative data* (2nd ed.). London: Sage Publications Inc.

Spearman, C. (1904). "*General intelligence", objectively determined and measured.* Retrieved September, 16, 2006 from http://psychclassics.yorku.ca/Spearman/chap1-4.htm

Stoneberg, B. D. (2004). *A study of gender-based and ethnic-based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests applying the* Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi-Square Test. Boise, ID: Idaho Department of Education.

Szabó, G. (2001). *The application of Item Response Theory in the construction of a language test item bank.* Unpublished doctoral dissertation. Pécsi Egyetem, Pécs.

Szabó, G. (2006). Anchors aweigh! An analysis of the impact of anchor item's

number and difficulty range on item difficulty calibrations. In M. Nikolov, & J. Horváth (Eds.), *University of Pécs Roundtable 2006: Empirical Studies in English Applied Linguistics* (pp.249-262). Pécs: Lingua Franca Csoport.

Takala, S. & Kaftandjieva, F. (2000). Test fairness: a DIF analysis o fan L2 vocabulary test. *Language Testing, 17(3),* 323-340.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 469-477.

Tyndall, B., & Kenyon, D. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing* (p. 9–57).Clevedon, England: Multilingual Matters.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Westport, Ct: Ablex Publishing.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One-parameter logistic model OPLM. Arnheim: Cito.

Viswanathan, M. (2005). Measurement error and research design. London: Sage Publications Ltd.

Weigle, S.C. (1994). Effects of training of raters of ESL compositions. *Language Testing, 11 (2)*, 197-223.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15 (2),* 263-287.

Weigle, S.C. (2000). Investigating rater/prompt interactions in writing assessment: qualitative and quantitative approaches. *Assessing Writing, 6(2)*, 145-178.

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

Weir, C. J. (2005). *Language testing and validation*. New York: Palgrave, Macmillan

Weitzman, E.A. (2000) Software and qualitative research, In N. K. Denzin, & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 803-820).

Wigglesworth, J. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10(3),* 305-335.

Wilson M.R. & Case H. (2000) An examination of variation in rater severity over time: a study of rater drift. In M. Wilson, & G. Engelhard, Jr. (Eds.) *Objective Measurement: Theory into practice: Vol. 5* (pp. 113-134). Stamford CT: Ablex Publishing.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46(1),* 35-51.

Wolfe, E. W., Moulder, B.C., & Myford, C.M. (2001). Detecting differential rater functioning over time using a Rasch Multi-Faceted rating scale model. *Journal of Applied Measurement, 2(3)*, 256-280.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.

Wright, B, & Linacre, J. 1994. Reasonable Mean-square Fit Values. *Rasch Measurement*
Transactions 8(3): 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). Differential item functioning. In *ACER ConQuest: Generalized item response modelling software* (pp. 75-84). Melbourne, Australia: Australian Council for Educational Research.

Appendices

Appendix A Glossary of Rasch terminology

bias	a factor of factors inherent within a test that systematically prevents access to valid estimates of candidate's ability.
discrepancy	one or more unexpected responses.
facet	in the analysis of test data, a measurable aspect of a performance or its setting which is hypothesized to have an impact on scores; a technical term in generalisability theory and multi-faceted Rasch measurement.
fit	the degree of match between the pattern of observed responses and the modelled expectation
fit statistic	a summary of the discrepancies between what is observed and what we expect to observe.
infit	a type of statistic used in item response theory analysis to indicate the extent of score variability in a given data set which remains after the extreme values have been removed
logit	the unit of measure used by Rasch for calibrating items and measuring persons. A log odds transformation of the probability of a correct response.
misfit	in Rasch analysis, a type of model data fit, reported in fir statistics for estimates of test item difficulty, candidate ability, rater severity and other facets of the assessment context and their interactions. Misfit indicates a lack of consistency in the score patterns associated with the facet concerned.
overfit	one type of poor fit, or failure of aspects of test score data to conform to the predictions of a data model, e.g. an IRT model. In analyses using probabilistic models, the model expects some variability from its expectations, within certain predicted limits. When this variability is significantly less than predicted, it is reported as overfit.
Rasch Model	a mathematical formula for the relationship between the probability of success (P) and the difference between an individual's ability (B) and an item's difficulty (D). P=exp(B-D)/(1+exp(B-D)) or log [P/(1-P)] = B - D
standardized z scores	a standardized z-score represents both the relative position of an individual score in a distribution as compared to the mean and the variation of scores in the distribution

Based on Davies et al. Dictionary of language testing and Wright, B.D. & Linacre J.M. (1985) Microscale Manual. Westport, Conn.: Mediax Interactive Technologies, Inc. retrieved from http://www.rasch.org/rmt/glossary.htm

Appendix B Sample FACETS analysis output

```
Facets for Windows Version No. 3.61.0  Copyright © 1987-2006, John M.
Linacre. All rights reserved.

2004_German writing#  ; Data specification
Facets = 3
Non-centered =  1
Positive =  1
Labels =
 1,Student (elements = 268)
 2,Rater (elements = 4)
 3,Ctiteria (elements = 4)
Model =?,#,?B,R6,1

; Output description
Arrange tables in order = MN
Bias/Interaction direction =ability ; leniency, easiness: higher score
= positive logit
Fair score = Mean
Heading lines in output data files = Y
Inter-rater coefficients reported for facet = 2
Omit unobserved elements = yes
Unexpected observations reported if standardized residual >= 3
Usort unexpected observations sort order = u
WHexact - Wilson-Hilferty standardization = Y

; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)

2004_German writing#  09-29-2006 13:34:07
Table 2. Data Summary Report.

Assigning models to "C:\Documents and
Settings\administrator\Dokumentumok\Eszter\PhD\Thesis\Facets
elemei\2004_01_German_specification#.txt"
Total lines in data file = 536
Total data lines = 536
Responses matched to model: ?,#,?B,R6,1 = 2136
    Total non-blank responses found = 2136
Number of missing-null observations = 8
Valid responses used for estimation = 2136

2004_German writing#  09-29-2006 13:34:07
Table 3. Iteration Report.

-------------------------------------------------------------
| Iteration      Max. Score Residual      Max. Logit Change |
|             Elements    %  Categories   Elements    Steps |
-------------------------------------------------------------
| PROX   1                                   .9888   2.5365 |
| PROX   2                                   .9883          |
| PROX   3                                   .9181          |
| PROX   4                                   .9194          |
| PROX   5                                   .6197          |
| JMLE   6   1273.9520 -61.5  -459.5621     -.9970   1.1097 |
| JMLE   7   -339.1441 -26.4  -317.0202     -.9647   1.1450 |
| JMLE   8   -112.0154  -9.5  -144.0313     -.2167   1.0788 |
| JMLE   9     73.5217 -15.5    53.9534      .4393   -.9756 |
```

JMLE 10	47.4713	-14.2	29.9543	.5432	.8916
JMLE 11	42.7284	-10.7	19.1786	.4715	-.7449
JMLE 12	48.0560	-8.7	17.7111	.3903	-.6376
JMLE 13	54.2426	-7.5	16.7515	.3379	-.3200
JMLE 14	44.7372	-6.6	14.2290	.2881	.2318
JMLE 15	38.1007	-5.8	12.5287	.2549	.2035

JMLE 380	.5277	.0	-.1415	-.0005	.0002
JMLE 381	.5244	.0	-.1407	-.0005	.0002
JMLE 382	.5236	.0	-.1398	-.0005	.0002
JMLE 383	.5184	.0	-.1389	-.0005	.0002
JMLE 384	.5159	.0	-.1381	-.0005	.0002
JMLE 385	.5128	.0	-.1372	-.0005	.0002
JMLE 386	.5109	.0	-.1364	-.0005	.0002
JMLE 387	.5051	.0	-.1354	-.0005	.0002
JMLE 388	.5021	.0	-.1346	-.0005	.0002
JMLE 389	.4993	.0	-.1339	-.0005	.0002

Subset connection O.K.

2004_German writing#
Table 4. Unexpected Responses - appears after Table 8.

2004_German writing#
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1*,2A,3A) Yardstick (columns,lines,low,high)= 0,2,-8,10

```
-------------------------------------------------------------------------------
-----
|Measr|+Student  |-Rater  |-Ctiteria                          | S.1 | S.2 | S.3 |
S.4 |
-------------------------------------------------------------------------------
-----
+  10 +          +        +                                   + (5) + (5) + (5) +
(5) +
|     | .        |        |                                   |     |     |     |
|
+   9 +          +        +                                   +     +     + --- +
+
|     | *        |        |                                   |     | --- |     |
|
+   8 +          +        +                                   +     +     +     +
--- +
|     | .        |        |                                   |     |     |     |
|
+   7 + **.      +        +                                   +     +     +  4  +
+
|     | *.       |        |                                   | --- |     |     |
|
+   6 + *.       +        +                                   +     +  4  +     +
4  +
|     | **.      |        |                                   |     |     |     |
|
+   5 + **       +        +                                   +     +     + --- +
+
|     | *        |        |                                   |  4  |     |     |
|
+   4 + ***.     +        +                                   +     +     +     +
--- +
|     | **       |        |                                   |     | --- |     |
|
+   3 + ****.    +        +                                   +     +     +     +
+
|     | *****    |        |                                   | --- |     |     |
|
+   2 + ****.    +        +                                   +     +     +  3  +
3  +
|     | ***.     |        |                                   |     |  3  |     |
|
+   1 + ******.  + Rater13+                                   +  3  +     +     +
+
|     | *****    | RaterG4| Style                             |     |     |     |
|
*   0 * *******. *        * Grammar           Task achievement *     *     *     *
--- *
|     | *****.   | Rater7 | Vocabulary                        | --- |     |     |
|
+  -1 + ******.  + RaterG2+                                   +     + --- + --- +
+
|     | ****.    |        |                                   |  2  |     |     |
|
+  -2 + ***.     +        +                                   +     +     +     +
+
|     | ****.    |        |                                   |     |  2  |  2  |
2  |
+  -3 + *.       +        +                                   + --- +     +     +
+
|     | **.      |        |                                   |     |     |     |
|
+  -4 + .        +        +                                   +  1  + --- +     +
+
|     | *.       |        |                                   |     |     | --- |
|
```

```
+  -5 + .        +        +                                  +     +     +     +
--- +
|     | *        |        |                                  | ---  |  1  |     |
|
+  -6 + .        +        +                                  +     +     +     +
1  +
|     |          |        |                                  |     |     |  1  |
|
+  -7 + .        +        +                                  +     + --- +     +
+
|     | .        |        |                                  |     |     |     |
--- |
+  -8 +          +        +                                  + (0) + (0) + (0) +
(0) +
-------------------------------------------------------------------------------------
-----
|Measr| * = 3    |-Rater  |-Ctiteria                         | S.1 | S.2 | S.3 |
S.4 |
-------------------------------------------------------------------------------------
-----
```

```
2004_German writing#
Table 6.1  Student Facet Summary.

Logit:
                        1   1 1 1 11  1  1  1 1
   1 1   2 3 11132143145342732097360 917339 2 115281 6 344122 7 2 3
11
 +---+---+-Q-+---+---+--S+---+---+---M---+---+---+---S---+---+---+---+Q--+---
+---+
-8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6   7   8   9
10

Infit MnSq:
 1      111211211
 071 17942313924186756868714441 33  11 12 1 2   21
1
 +----S---------M--------S-----+---Q----------+--------------+--------
------+
 0              1              2              3              4
5

Outfit MnSq:
 1      11111111      1 1
 071 59804828983994766160551312 33 1 11 22 1 1 1  11   1
1
 +----S---------M---------S----+----Q---------+--------------+--------
------+
 0              1              2              3              4
5

Infit ZStd:
                            11 1 11111    1 1
              1 24  92  173825829898419766282845251111211 1     1
 +------+------+---Q--+----S-+-----M+------+S-----+-Q----+------+-----
-+
-5     -4     -3     -2     -1      0      1      2      3      4
5

Outfit ZStd:
                      1      111 1111111   111
             1   33  1  44565408365732068810067351 3 1122 11     1
 +------+------+---Q--+----S-+-----M+------+S-----+-Q----+------+-----
-+
-5     -4     -3     -2     -1      0      1      2      3      4
5


2004_German writing#
Table 6.2  Rater Facet Summary.

Logit:
                             11     11
 +---+---+---+---+---+---+---+Q--+S--M--S+--Q+---+---+---+---+---+---+---+---
+---+
-8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6   7   8   9
10

Infit MnSq:
              1 21
 +-----------Q-SMS-Q-----------+--------------+--------------+--------
------+
```

```
 0              1              2              3              4
5

Outfit MnSq:
             1  21
 +-----------QS-MS-Q-----------+--------------+--------------+--------
------+
 0              1              2              3              4
5

Infit ZStd:
                 1                    1  1     1
 +------+-----Q+------+-S----+-----M+------+--S---+-----Q+------+-----
-+
-5     -4     -3     -2     -1      0      1      2      3      4
5

Outfit ZStd:
                 1                   1    1      1
 +------+--Q---+------+S-----+----M-+------+--S---+------Q------+-----
-+
-5     -4     -3     -2     -1      0      1      2      3      4
5


2004_German writing#
Table 6.3  Ctiteria Facet Summary.

Logit:
                                1 2 1
 +---+---+---+---+---+---+---+QS-M-SQ+---+---+---+---+---+---+---+---
+---+
-8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6   7   8   9
10

Infit MnSq:
            1  1  11
 +---------Q--S-M--S-Q---------+--------------+--------------+--------
------+
 0              1              2              3              4
5

Outfit MnSq:
            1  1  2
 +---------Q-S--M-S--Q---------+--------------+--------------+--------
------+
 0              1              2              3              4
5

Infit ZStd:
      1                         1               1      1
 +------+------+S-----+------+-----M+------+------+---S--+------+-----
-+
-5     -4     -3     -2     -1      0      1      2      3      4
5

Outfit ZStd:
     1                          1              1    1
 +------+------+S-----+------+----M-+------+------+-S----+------+-----
-Q
```


-5 -4 -3 -2 -1 0 1 2 3 4
5

2004_German writing#
Table 7.2.1 Rater Measurement Report (arranged by MN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	N Rater
1613	532	3.0	2.97	-.94	.09	.85	-2.6	.82	-2.8	1.16	68.4	56.1	4 RaterG2
1179	400	2.9	2.96	-.73	.11	1.01	.1	.99	-.1	.98	72.8	58.0	3 Rater7
1368	536	2.6	2.72	.74	.07	1.03	.5	1.03	.5	.97	50.7	48.0	1 RaterG4
1734	668	2.6	2.75	.93	.07	1.08	1.4	1.09	1.5	.91	51.6	48.5	2 Rater13
1473.5	534.0	2.8	2.85	.00	.09	.99	-.1	.98	-.2				Mean (Count: 4)
215.2	94.8	.2	.12	.84	.01	.09	1.5	.10	1.6				S.D. (Populn)
248.4	109.4	.2	.13	.97	.02	.10	1.8	.12	1.9				S.D. (Sample)

Model, Populn: RMSE .09 Adj (True) S.D. .84 Separation 9.58 Reliability (not inter-rater) .99
Model, Sample: RMSE .09 Adj (True) S.D. .97 Separation 11.07 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 386.8 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 3.0 d.f.: 2 significance (probability): .22
Inter-Rater agreement opportunities: 1068 Exact agreements: 636 = 59.6% Expected: 555.7 = 52.0%

2004_German writing#
Table 7.3.1 Ctiteria Measurement Report (arranged by MN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Ctiteria
1549	534	2.9	3.01	-.53	.08	.75	-4.3	.73	-4.4	1.25	2	Vocabulary
1489	534	2.8	2.92	-.11	.08	1.18	2.7	1.16	2.2	.83	4	Grammar
1460	534	2.7	2.87	.10	.08	1.11	1.7	1.11	1.5	.89	1	Task achievement
1396	534	2.6	2.77	.54	.08	.96	-.6	.96	-.5	1.03	3	Style
1473.5	534.0	2.8	2.89	.00	.08	1.00	-.1	.99	-.3			Mean (Count: 4)
55.1	.0	.1	.09	.38	.00	.16	2.7	.17	2.6			S.D. (Populn)
63.6	.0	.1	.10	.44	.00	.19	3.1	.19	3.0			S.D. (Sample)

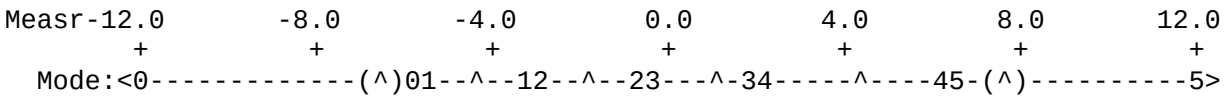
Model, Populn: RMSE .08 Adj (True) S.D. .37 Separation 4.47 Reliability .95
Model, Sample: RMSE .08 Adj (True) S.D. .43 Separation 5.20 Reliability .96
Model, Fixed (all same) chi-square: 84.0 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 2.9 d.f.: 2 significance (probability): .23

2004_German writing# 09-29-2006 13:34:07
Table 8.1 Category Statistics.

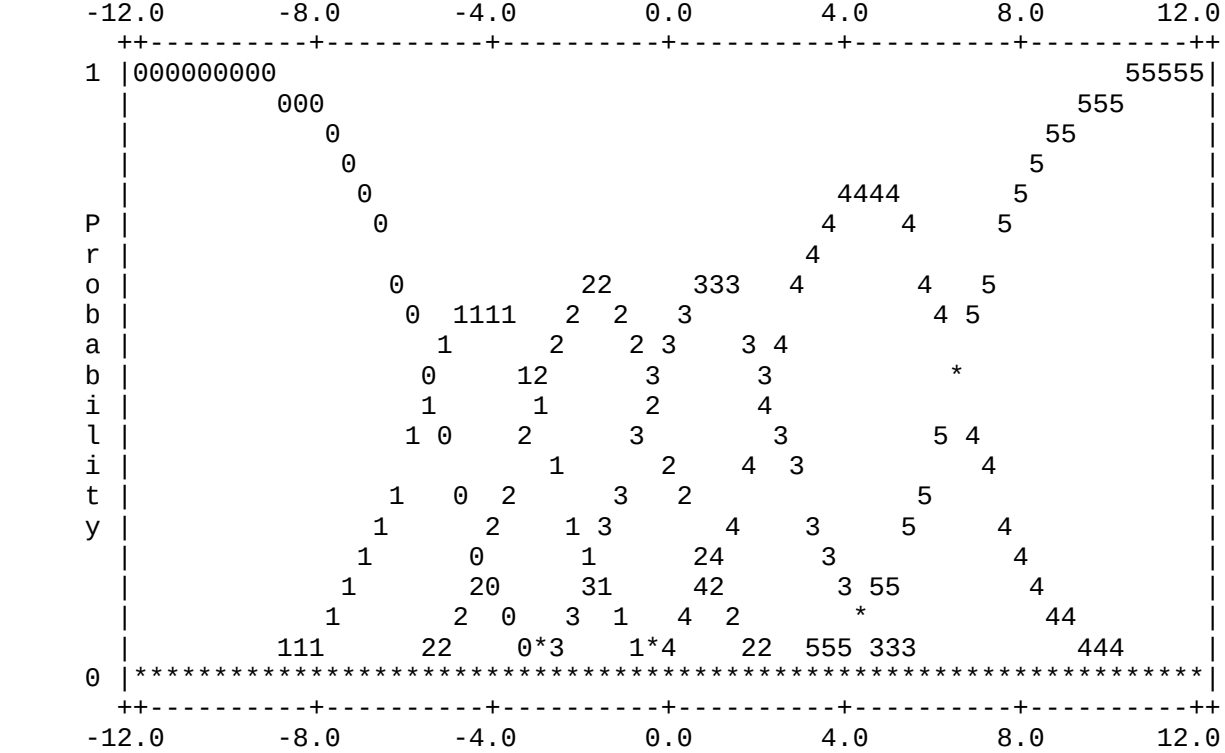
 Model = ?,1,?B,R6   RaterG4

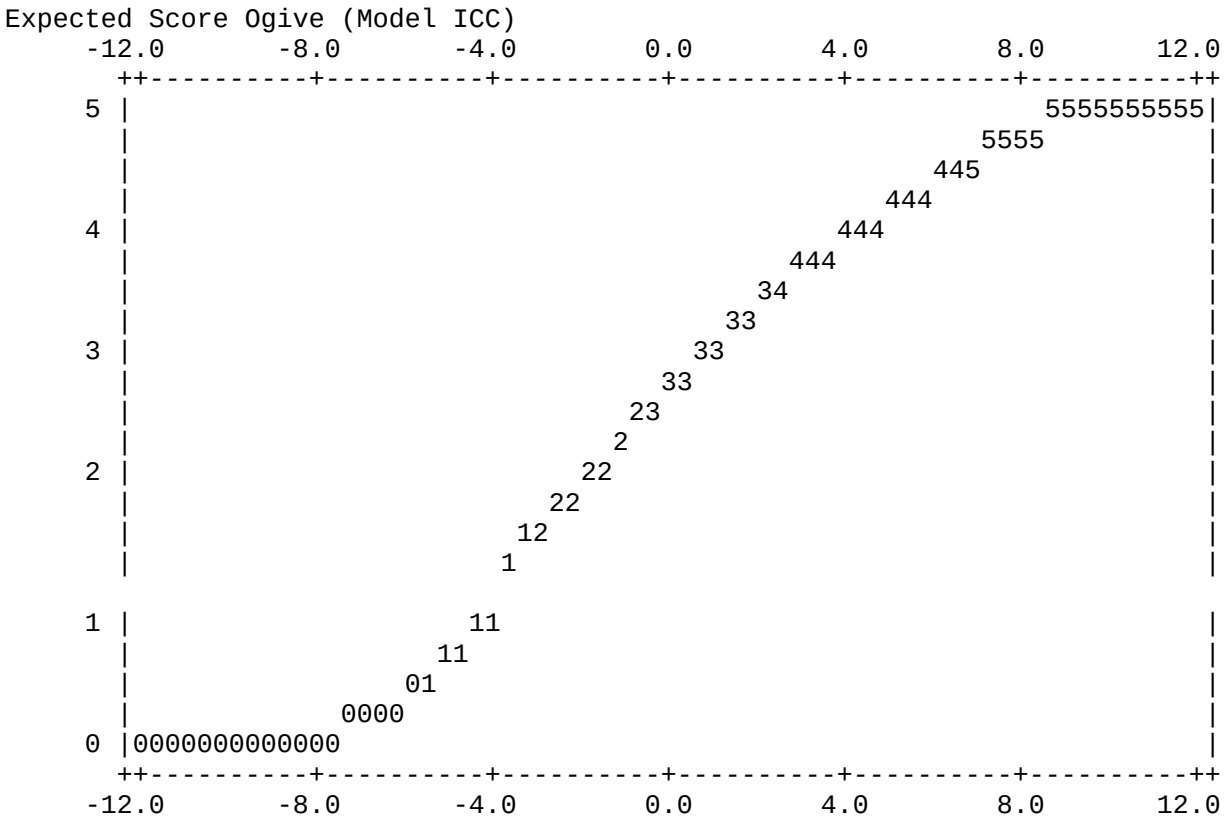
DATA			QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
0	16	3%	-4.44	-4.50	1.1			(-6.55)		low	low	100%
1	76	14%	-3.16	-3.09	.8	-5.42	.30	-4.23	-5.60	-5.42	-5.50	62%
2	171	32%	-1.27	-1.36	1.1	-3.03	.16	-1.70	-2.99	-3.03	-3.01	64%
3	160	30%	.43	.46	1.1	-.42	.13	.94	-.39	-.42	-.41	66%
4	95	18%	3.25	3.34	1.1	2.28	.17	4.42	2.40	2.28	2.32	81%
5	18	3%	6.36	6.18	.9	6.58	.31	(7.67)	6.61	6.58	6.59	100%
		100%						(Mean)		(Modal)	(Median)	

Scale structure



Probability Curves



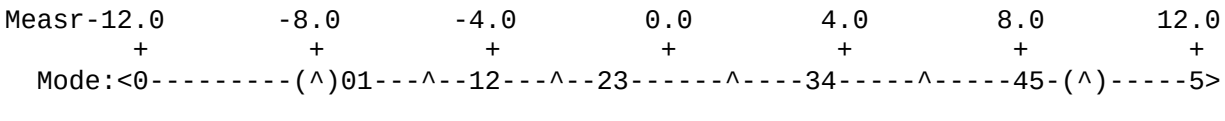

Expected Score Ogive (Model ICC)
-12.0 -8.0 -4.0 0.0 4.0 8.0 12.0
5
4
3
2
1
0
5555555555
5555
445
444
444
444
34
33
33
33
23
2
22
22
12
1
11
11
01
0000
0000000000000
-12.0 -8.0 -4.0 0.0 4.0 8.0 12.0

2004_German writing#
Table 8.2 Category Statistics.

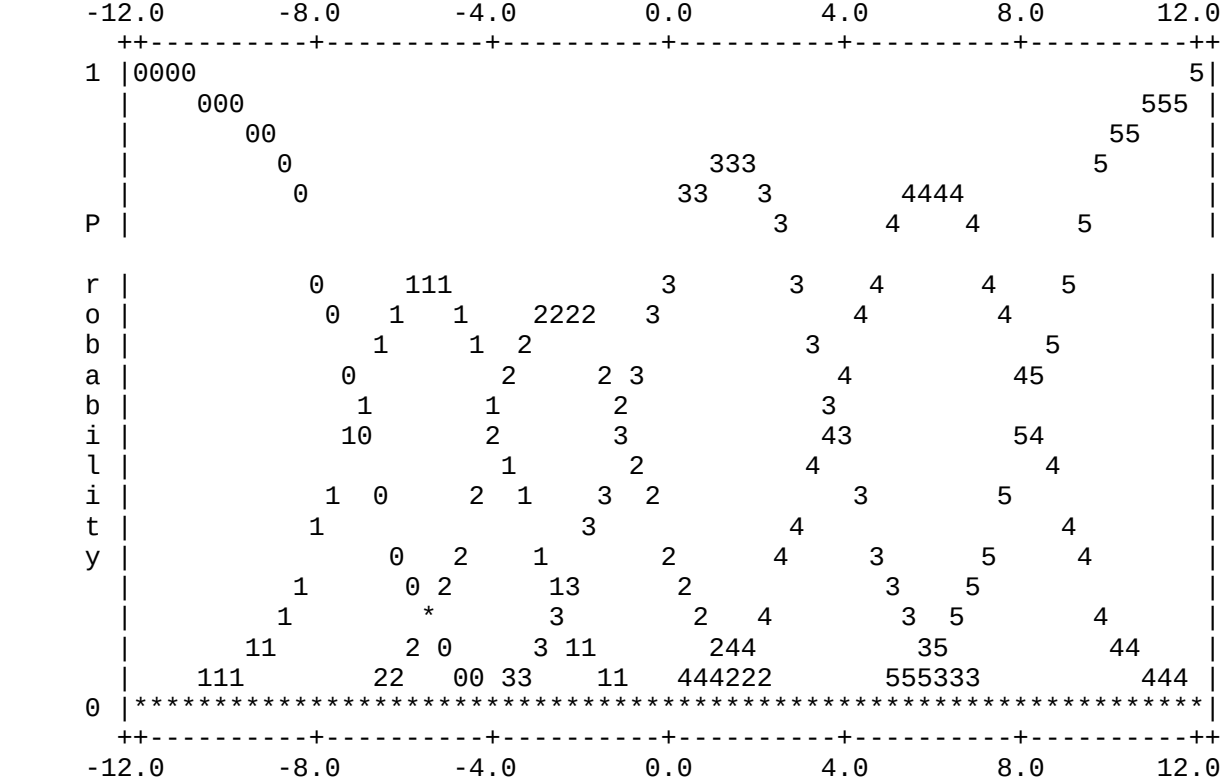
Model = ?,2,?B,R6 Rater13

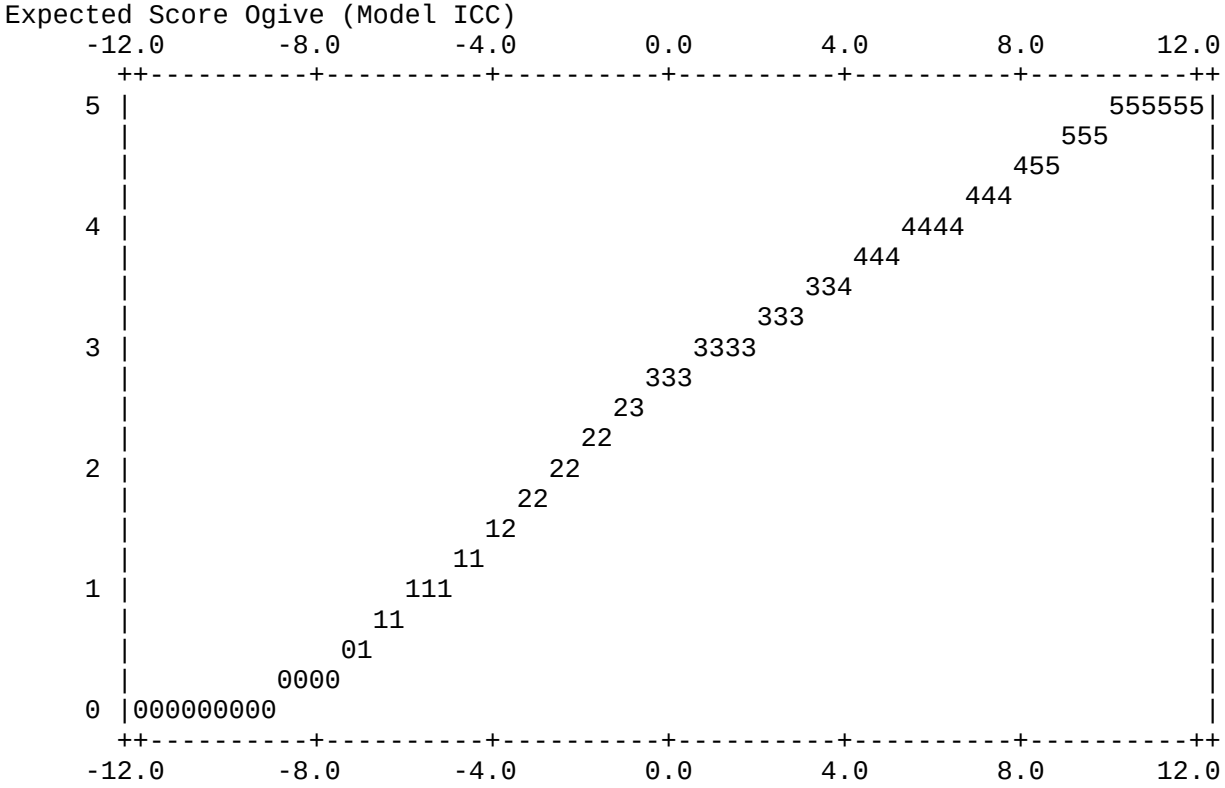
DATA			QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
0	8	1%	-5.19	-6.37	2.0			(-8.15)		low	low	100%
1	70	10%	-3.84	-3.75	1.0	-7.06	.45	-5.45	-7.14	-7.06	-7.09	71%
2	208	31%	-1.77	-1.78	1.1	-3.86	.16	-2.42	-3.88	-3.86	-3.87	67%
3	289	43%	.58	.57	1.0	-1.03	.11	1.31	-.89	-1.03	-.98	84%
4	84	13%	4.43	4.49	1.1	3.72	.18	5.94	3.70	3.72	3.70	83%
5	9	1%	100%	6.83	7.20	1.1	8.23	.40	(9.30)	8.24	8.23	8.23	100%
								(Mean)		(Modal)	(Median)	

Scale structure



Probability Curves



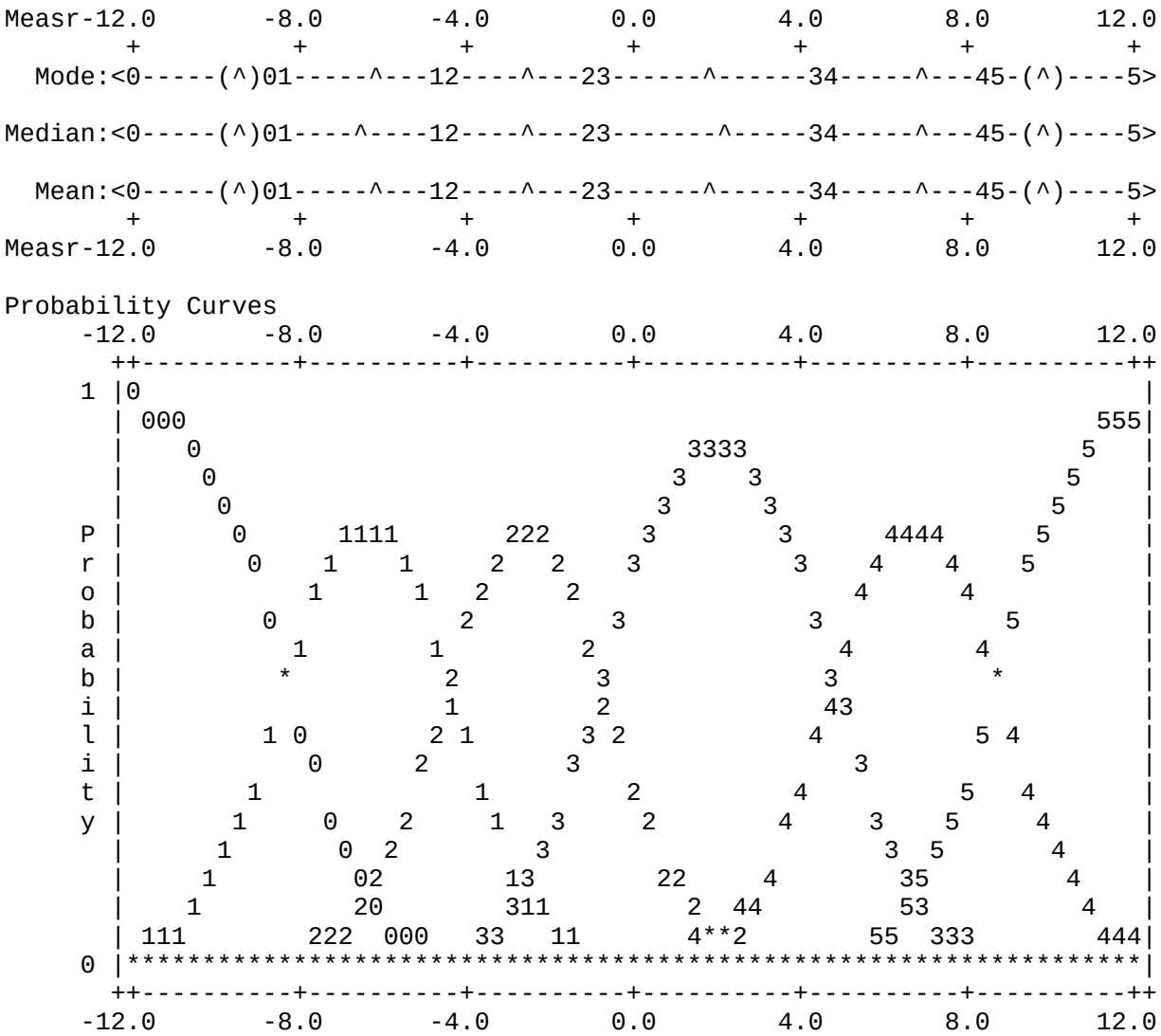

Expected Score Ogive (Model ICC)
-12.0 -8.0 -4.0 0.0 4.0 8.0 12.0
5
4
3
2
1
0
555555
555
455
444
4444
444
334
333
3333
333
23
22
22
22
12
11
111
11
01
0000
000000000

2004_German writing#
Table 8.3 Category Statistics.

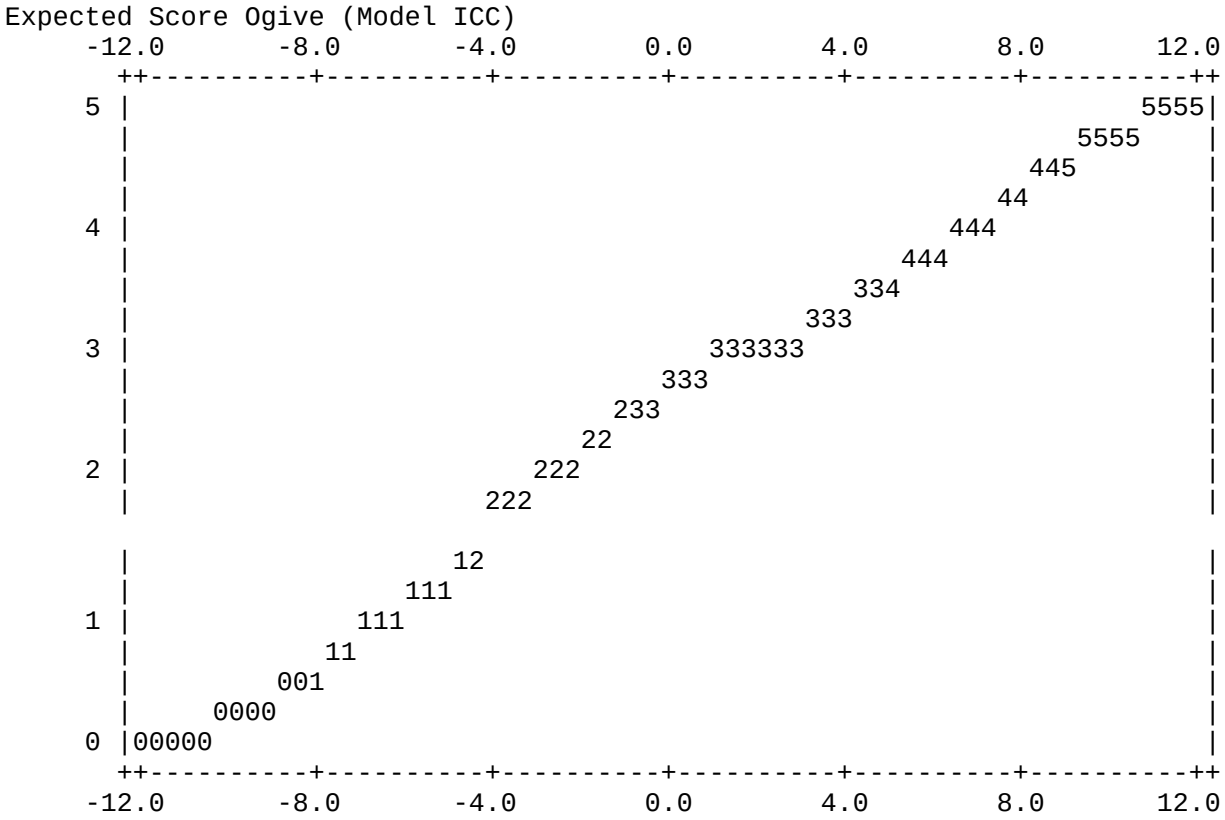
Model = ?,3,?B,R6 Rater7

DATA			QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
0	1	0%	-6.45	-5.73	.8			(-9.38)		low	low	100%
1	22	6%	-4.56	-4.54	.8	-8.30	1.04	-6.34	-8.33	-8.30	-8.31	78%
2	80	20%	-1.56	-1.52	.9	-4.42	.31	-2.59	-4.42	-4.42	-4.42	75%
3	207	52%	2.00	1.93	1.1	-.82	.18	1.94	-.73	-.82	-.79	89%
4	74	19%	5.49	5.68	1.1	4.84	.19	6.74	4.78	4.84	4.81	77%
5	16	4%	8.48	8.29	.8	8.69	.33	(9.79)	8.74	8.69	8.70	100%
		100%						(Mean)		(Modal)	(Median)	

Scale structure



Probability Curves


Expected Score Ogive (Model ICC)
-12.0 -8.0 -4.0 0.0 4.0 8.0 12.0
5
4
3
2
1
0
00000
0000
001
11
111
12
111
222
222
233
22
333
333333
333
334
444
444
445
44
5555
5555

2004_German writing#
Table 8.4 Category Statistics.

 Model = ?,4,?B,R6   RaterG2

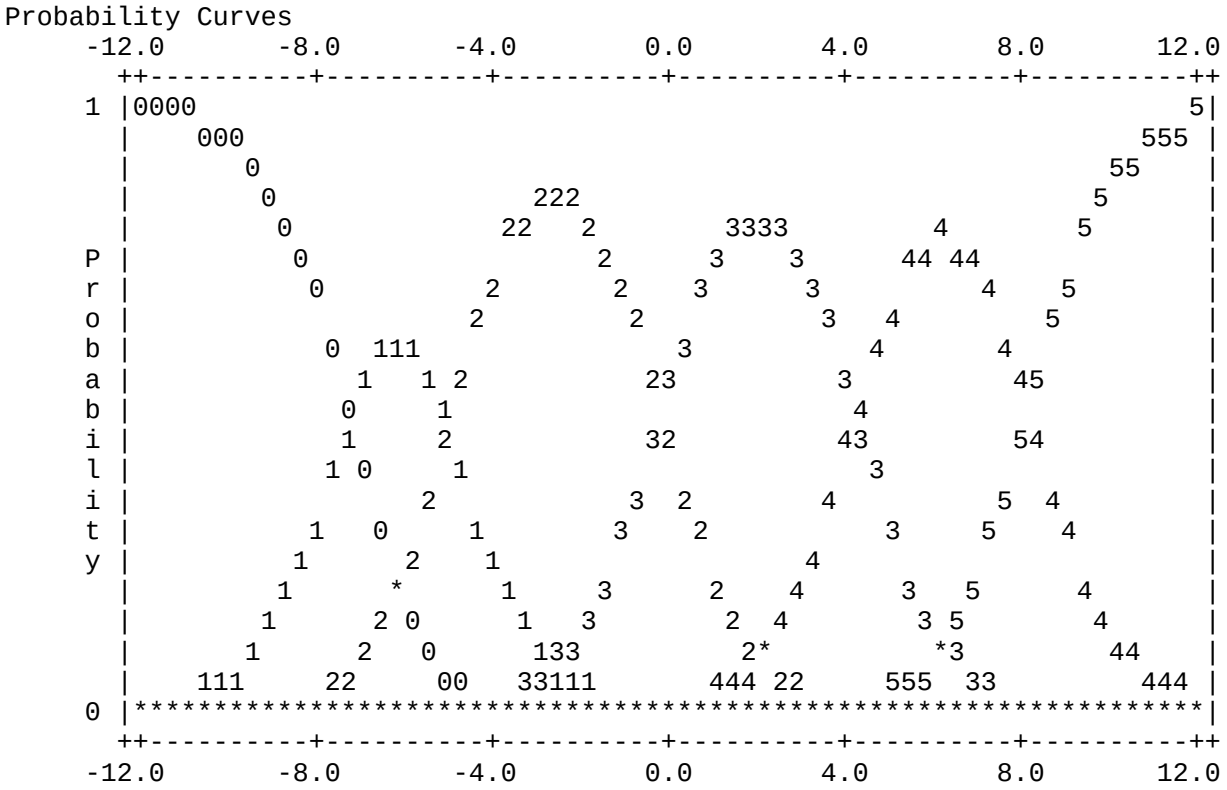
DATA			QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
0	3	1%	-6.72	-6.09	.5			(-8.36)		low	low	100%
1	17	3%	-4.77	-4.67	.8	-7.22	.66	-6.12	-7.44	-7.22	-7.31	60%
2	128	24%	-1.27	-1.19	.8	-5.05	.32	-2.56	-4.82	-5.05	-4.96	85%
3	231	43%	1.94	1.96	.9	-.18	.15	2.01	-.17	-.18	-.18	82%
4	118	22%	5.41	5.29	.8	4.27	.16	6.21	4.23	4.27	4.25	78%
5	35	7%	8.52	8.39	.9	8.19	.25	(9.27)	8.23	8.19	8.20	100%
		100%						(Mean)		(Modal)	(Median)	

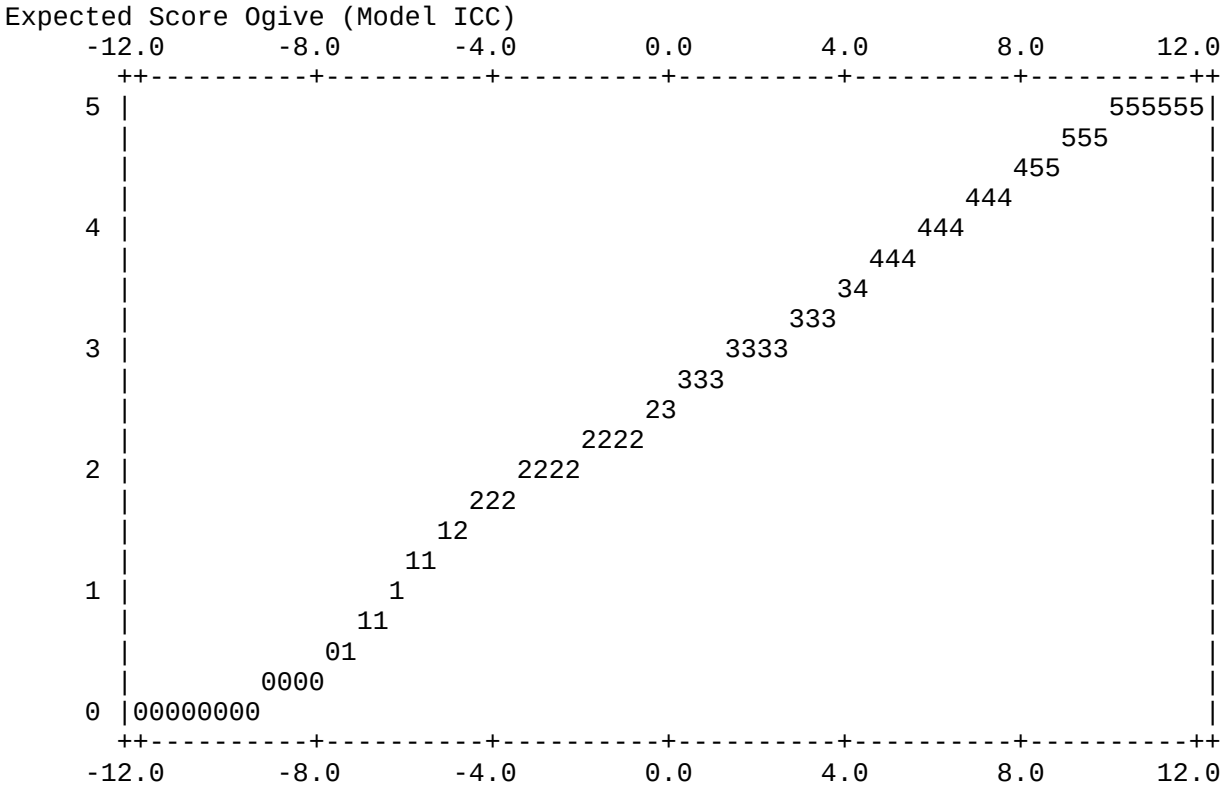
Scale structure

```
Measr-12.0       -8.0       -4.0        0.0        4.0        8.0       12.0
        +          +          +          +          +          +          +
  Mode:<0--------(^)01--^-12------^----23------^----34----^----45(^)------5>

Median:<0--------(^)01--^-12------^----23------^----34----^----45(^)------5>

  Mean:<0--------(^)01--^--12-----^-----23-----^----34----^----45-(^)-----5>
        +          +          +          +          +          +          +
Measr-12.0       -8.0       -4.0        0.0        4.0        8.0       12.0
```


Probability Curves
-12.0
-8.0
-4.0
0.0
4.0
8.0
12.0
Probability
1
0


Expected Score Ogive (Model ICC)
-12.0 -8.0 -4.0 0.0 4.0 8.0 12.0
5
4
3
2
1
0
555555
555
455
444
444
444
34
333
3333
333
23
2222
2222
222
12
11
1
11
01
0000
00000000

2004_German writing#
Table 4.1 Unexpected Responses (10 residuals sorted by u).

Cat	Step	Exp.	Resd	StRes	Num	Studen N	Rater	N	Ctiteria
5	5	3.3	1.7	3	13	040112 2	Rater13	4	Grammar
1	1	2.7	-1.7	-3	22	040121 2	Rater13	1	Task achievement
0	0	1.8	-1.8	-3	94	040192 2	Rater13	4	Grammar
4	4	3.0	1.0	3	142	040240 3	Rater7	3	Style
4	4	3.0	1.0	3	164	040262 3	Rater7	4	Grammar
2	2	3.1	-1.1	-3	170	040268 3	Rater7	4	Grammar
1	1	2.8	-1.8	-3	214	040312 4	RaterG2	1	Task achievement
4	4	1.9	2.1	3	259	040357 1	RaterG4	4	Grammar
0	0	1.9	-1.9	-3	259	040357 2	Rater13	3	Style
0	0	2.2	-2.2	-3	265	040363 2	Rater13	4	Grammar
Cat	Step	Exp.	Resd	StRes	Num	Studen N	Rater	N	Ctiteria

Appendix C Tasks used in Study 2

The English task

WRITE ON THE ANSWER SHEET.
YOU CAN USE THIS SHEET FOR YOUR DRAFT.

You are Emese / Endre Kisházy, the Product Manager of a Travel Agency called Hungária
(1085, Budapest, József krt. 43). You have just received a letter from a Chinese tour operator (Sunrise Travel, Mr Zhu He-Ling, 28 Guomenwa Avenue, Chaoyanggu, Beijing, 100740) who would like to have holidays organized for Chinese groups in Hungary. He makes enquiries about:

- Recommended destinations inside the country
- Available types of accommodation
- Possible 10 day itineraries for the groups
- Trial costing

Answer this letter in about 150-200 words. Include the following points:

- Thank him, indicate that such a cooperation is very important for you
- Ask for more detailed information on potential guests
- Give a kind of sample itinerary, mention a few recommended destinations in the country
- Enclose a brochure with itineraries and prices of tours your agency organizes in Hungary for foreign visitors in 2004

The German Task

FÜR DIE LÖSUNG DER AUFGABE BENUTZEN SIE DAS LÖSUNGSBLATT!

Sie sind:	Gabriella / Gergely Baranyai, Verkaufsmanager in der Építők Kft. H-1145 Budapest, Gabona u. 3.
Sie schreiben an:	Tondach Magyarország Rt. (Hersteller von Tondachziegel in Mittel-/Osteuropa H-1124 Budapest Németvölgyi út 100.
Sie schreiben:	eine **Bewerbung** um die Position Verkaufsleiter für Ostungarn
Bezug:	Stellenanzeige in der HVG
Aufgabe des Kandidaten:	- Leitung und Ausbau der Verkaufsmannschaft - Kundenbetreuung, -gewinnung - Projektarbeit - Kontaktpflege (Architektenbüros, öffentliche Ämter, Denkmalschutz, Kirche, usw.) - Vertriebskonzeptentwicklung in der Region
Voraussetzungen:	- kaufm./techn. Ausbildung - Sprach- und EDV - Kenntnisse
Länge:	**150-200 Wörter**

Appendix D The rating scale validated in the study

KÖZÉPFOK

SZAKMAI SZÖVEG ÍRÁSA

Az értékelési szempontok sávleírása

	feladatmegoldás	**szókincs/frazeológia**	**szövegkezelési technika**	**nyelvhelyesség/helyesírás**
5	a vizsgázó a feladatban megjelölt szövegfajtának megfelelő szöveget készít, melyben a tartalmi pontok logikusan, kellőrészletességgel szerepelnek	a vizsgázó a tartalomnak megfelelő szavakat, szakkifejezéseket és fordulatokat használ	a vizsgázó a műfajnak megfelelő stílusú, szerkezetű és formájú szöveget alkot	a vizsgázó szövegét a szövegrészek logikus kapcsolása, a nyelvi szerkezetek adekvát használata jellemzi, a nyelvi és helyesírási hibák mennyisége elenyésző
4	a feladatban megjelölt szövegfajtának megfelelő szöveget készít, melyben szerepelnek a szükséges közlendők	a tartalomnak többnyire megfelelő szavakat, szakkifejezéseket és fordulatokat használ	a műfajnak megfelelő stílusú és formájú szöveget alkot némi szerkesztési hiányossággal vagy aránytalansággal	szövegét jól megformált mondatok, megfelelő nyelvi szerkezetek jellemzik, de több nyelvi és helyesírási hibát vét
3	a feladatban megjelölt szövegfajtának megfelelő szöveget készít, amelyben a szükséges közlendők nagyrészt megvannak, de szükségtelen pontok is szerepelnek	helyenként nem a megfelelő szavakat, szakkifejezéseket és fordulatokat használja	a műfaj szempontjából elfogadható szöveget készít, de annak stílusa egyenetlen, és a szövegben szerkesztési hibák is vannak	többnyire egyszerű mondatokból álló szövegében nyelvi hibák és ismétlődő helyesírási hibák vannak az értelem torzítása nélkül
2	nem a tartalomnak megfelelő szövegfajtát készít, a tartalom logikátlan és hiányos	alig használ megfelelő szakkifejezéseket, fordulatokat	szövegében a műfajra jellemző sajátságok csak elemekben fedezhetők fel	szövegét rosszul formált, hiányos, töredékes mondatok, nyelvi és helyesírási hibák jellemzik
1	nem készít önálló szöveget, a megadott szempontokat ismétli, reprodukálja töredezetten	szóhasználata több helyütt is félreérthető, nem vagy véletlenszerűen használ megfelelő szakkifejezéseket, fordulatokat	műfajilag nem megfelelő stílusú és terjedelmű szöveget készít	mondatai, mondattöredékei nem alkotnak szöveget, súlyos nyelvi hibák jellemzik írását
0	másról ír, vagy nem ír semmit	szóhasználata érthetetlen	inkoherens, töredékekből álló szöveget készít, vagy nem ír semmit	mondatai, mondattöredékei értelmezhetetlenek

Appendix E Sample English scripts for marking in Study 2

043532

Hungária
Travel Agency
József krt. 43.
Budapest
1085

Zhu He-Ling
Tour operator
28 Guomenwa Avenue
Chaoyanggu, Beijing 100740 30/04/2004

Dear Mr. He-Ling,

I have just received your letter, I am writing to you concerning your request about holidays organized for Chinese groups in Hungary.

First of all I would like to thank you for choosing our travel agency. Our cooperation is very important for us because we are specialized at organizing holidays for Chinese groups.

Be so kind and let me know please more information about the groups, for examples the average age of the participants and about the periods they should arrive.

I enclose a brochure for you which contains some itineraries and prices of tours organized by our agency for foreign visitors in 2004.

I ~~migh~~ suggest you some interesting destinations. First of all our capital, Budapest and Lake Balaton especially in the high season (summer).
Another interesting tour could be the Great Plain of Hungary. There a sample itinerary could be the following: Bugac – Debrecen – Kecskemét.

About the accommodation I would suggest to stay in private houses because there ~~you~~ the guests can enjoy the natural environment, except Budapest *.
If you have any question do not hesitate to contact us. We are at your disposal in everything.

Yours sincerely,
Emese Kisházy
Product Manager

*where my offer is to stay in hotel.

(227 words)

043537

Hungária Travel
Agency
Emese Kisházy
Product Manager
József krt. 43.
1085 Budapest
HUNGARY

Sunrise Travel
Mr Zhu He-Ling
Tour operator
28 Guomenwa Avenue
Chaoyanggu, Beijing
CHINA

30 April 2004

Dear Mr Zhu,

I thank you for your letter of 27 April.
I would like to express that Hungária Travel Agency is glad that we cooperate.

I will be pleased if you tell me more information about your guests who are going to visit (to) Hungary (e.g. sex, age, range of interests).
Our country has plenty of sights in every respect. Budapest is not only the capital of the country but the capital city of the thermal spas too. There are more than 20 baths which are very famous because its water cure a lot of desease.
Hungary has a lot of wine region.
The most famous is the Tokaj wine region where tourists can taste the popular nectar called 'Aszú'!

In Hungary we can visit a lot of castle (e.g. in Budapest, Eger, Gödöllő).
These are worth seeing.

I enclose the newest brochure which contains all the tours of us in 2004 (of course with prices).

I am looking forward to hearing from you soon.

Enclosure: 1

Yours sincerely,

Emese Kisházy
Emese Kisházy
Product Manager

043573

Hungária Travel Agency
43. József krt.
1085 Budapest
Hungary

30 april 2004

Sunrise Travel
28Guomenwa Avenue
Chaoyanggu,
Beijing
100740

Dear Mr Zhu He-Ling

Thank you very much for your recent letter enquiring about holiday in Hungary.
We are delighted to give you information about our country, tourism destinations and facilities.
We believe our cooperation will be successful

Hungary offers a wide range of facility~~ies~~ for visitors such as sightseeing tours, holiday at Lake Balaton or in the mountains.
Please let me know more details about your potential guest in order to give you the most suitable offer.
I am enclosing a brochure about ~~our~~ the tours we organise in Hungary for foreign visitors. As you will see, the most popular destination is the capital city called Budapest. The city offers is rich in historic monuments and offers great cultural and entertaining facilities. We suggest spending 2 or 3 days in the capital city. (The) Lake Balaton is also a very popular destination especially in the summer where visitors find great opportunity for recreation.
The Hortobágy or Great Plain is a symbol of Hungary. The destination offers the typical Hungarian restaurant so called "csárda".

Hungary offers a wide range of accommodation facilities from five-star luxurious hotels to traditional guest houses.

Please find more information about the destinations and accommodations in brochure. The brochure also includes the prices for this season.

We hope to have the pleasure of welcoming your guest or groups in our country.

Yours sincerely

Emese Kisházy
Product Manager

(250 words)

Appendix F Sample German scripts for marking in Study 2

044322

Tondach Magyarország Rt:

(Hersteller von Tondachziegel
in Mittel / Osteuropa)
H-1124 Budapest
Németvölgyi út 100.

Gabriella / Gergely Baranyai
Verkaufsmanager
In der Építõk Kft
H-1145 Budapest,
Gabona u. 3.

Bewerbung

Sehr geehrten Damen und Herren,

ich habe in der HVG eine Stellenanzeige gelesen.
Ich bin 24 Jahre alt. Ich möchte mich zur Arbeit melden.
Ich bin Verkaufsmanager. Ich habe ein Diplom, und ich arbeitete schon als Verkaufsleiter. Ich kann deutsch, englisch sprechen. Ich kann EDV-Kenntnisse auch.
- Meine Aufgabe wäre: Leitung und Ausbau der Verkaufsmannschaft,
- Kundenbetreuung –gewinnung
  Projektarbeit
- Kontaktpflege (Architektenbüros öffentliche Ämter)
Denkmalschutz, Kirchen,)
- Vertriebskonzeptentwicklung in der Region

Ich hoffe, dass ich diese Position als Verkaufsleiter gut bin. Faxen Sie mich bitte, wenn diese Position frei ist.
Ich möchte wissen, dass die Zahlungsleistung wie ist? Werde ich Tantieme bekommen?
Bitte schreiben oder faxen Sie von der Firma?
Bitte schreiben Sie von den Dienstleistungen? Wenn ich bei Ihrer Firma dort arbeiten werde, werde ich ein Auto bekommen?
Ich möchte mein Diplom und mein Lebenslauf und meine Sprachprüfung eine Schrift zu den Melden beibinden.

Mit freundlichen Grüssen:
Gabriella/Gergely Baranyai

Anlage2

044344
Gabriella Baranyai
H-1145 Budapest
Gabona u. 3.

Tondach Magyarország RT.
H-1124 Budapest
Németvölgyi út 100.

Budapest, 30. Április
2004

Sehr geehrte Damen /und Herren
ich habe Ihre Stellenanzeige in der HVG, in der Sie einen Verkaufsleiter besuchen.
Ich arbeite jetzt bei dem Építők KFT, als Verkaufsmanagerin. Das ist meine erste Arbeitsstelle. Ich bin mit den Geschäftspartnern der Firma im Kontakt. Ich muss verschiedene Wettbewerbe vorbereiten, und die zusammensammeln. Meine Aufgabe ist noch die Abwicklung der Bestellung. Meine Arbeit ist sehr interessant, aber ich möchte in einer höcheren Position arbeiten. Deshalb habe ich mich um die Position Verkaufsleiter bei Ihrer Firma bewerben. Ich bin anpassungsfähig, kann ich in einer Gruppe sehr gut arbeiten. Meine Kommunikationsfähigkeit ist sehr gut, nehme ich mit den Leuten einfach auf. Ich bin noch kreativ. Ich möchte an dem Erfolg der Firma teilnehmen, und mit meiner Arbeit in der Entwicklung helfen.
Ich möchte mit meiner Kenntnisse und Erfahrung die Ihre Firma helfen.
Ich habe Kaufman Ausbildung, die an der Wirtschaftliche Hochschule erwerben habe.
Meine Sprachkenntnisse ist von Deutsch und Englisch sehr gut, ich muss oft deutsch oder englisch sprechen.
Ich schließe zu meinem Brief mein Lebenslauf bei, in dem meine Ausbildung, Kenntnisse und Erfahrung angeführt werden.

Ich bedanke uns schon im Voraus, für die baldmöglichste Antwort.

Mit herzlichen Grüß
Gabriella Baranyai

044411

Gabriella Baranyai
Verkaufsmanagerin der Építõk Kft.
Gabona u. 3.
H-1145 Budapest

Tondach Magyarország Rt.

Németvölgyi út 100.
H-1124 Budapest

Bewerbung 05. 05. 2004

Sehr geehrte Damen und Herren,

ich habe Ihre Anzeige in der Zeitschrift HVG vom 3. Mai gelesen, und Ihr Stellenangebot hat mein Interesse geweckt. Ich möchte mich um die Position Verkaufsleiter für Ostungarn bewerben.

Ich interessiere mich für diese Position sehr, weil die von Ihnen in der Anzeige angeführten Aufgaben für mich neue Herausforderungen bedeuten würden, und ich könnte neben meinen Verkaufserfahrungen auch meine Marketingkenntnisse ausnutzen.

Ich habe mein Diplom an der Hochschule für Außenhandel gemacht, und ich arbeite seit 2000 auf dem Gebiet des Verkaufs bei der Építõk Kft. Ich spreche Deutsch und Englisch fließend, und ich lerne jetz auch Spanisch. Ich besitze EDV Kenntnisse, ich habe die ECDL Prüfung bestanden.

Ich habe Organisationstalent und ich kann mit anderen Menschen den Kontakt schnell aufnehemen, deshalb könnte ich die Kundenbetreuung sehr gut organisieren. Ich bin auf dem laufenden auf dem Gebiet der Bauindustrie und habe Kontakte zu den öffentlichen Ämtern.

Ich würde mich freuen, wenn Sie mir eine Gelegenheit zum persönlichen Zusammentreffen gewähren würden.
Mit freundlichen Grüßen:
Gabriella Baranyai
Gabriella Baranyai

Anlage:

1. Diplomduplikat
2. ECDL Zeugnisduplikat

Appendix G Interview protocol for the raters in Study 2

Interjú kérdések

Leniency/Severity/Generosity

Milyen értékelőnek tartod magad?

A pontszámok egyeztetésénél általában te vagy a szigorúbb vagy az enyhébb értékelő?

A pontszámok egyeztetésénél inkább meg akarod győzni a javítótársadat vagy inkább a két pontszám közötti értékben egyeztek meg?

Extremism/Central tendency

Mikor érzed indokoltnak a 0 pontot/maximális pontot?
(Ilyenkor adsz-e ilyen pontszámokat?)
Milyen esetekben adsz 0 pontot/maximális pontot?
Melyik szélső érték fordul elő gyakrabban az értékelésed során?

Halo/Carry over effect

Melyik értékelési szempontot találod a legfontosabbnak?
Melyik értékelési szempontot találod a legkevésbé fontosnak?
Milyen összefüggést látsz az értékelési szempontok között? (pl. ha ez a szempont x pont, akkor az a szempont nem lehet annál több.)

Instability

Melyek azok a tényezők, amelyek befolyásolják az értékítéletedet? (javítópár, javított dolgozatok mennyisége, feladat - saját, kedvelt szakágazat stb.-, rendelkezésre álló idő, egyéb)
Mitől értékelsz szigorúbban?
Mitől értékelsz enyhébben?
Mennyire tartod magad konzekvensnek egy értékelési perióduson belül? és
Mennyire tartod magad konzekvensnek több értékelési periódust figyelembe véve?

Kérlek, tedd sorrendbe az értékelési szempontokat az alábbi két tényező szerint. 1-gyel jelöld a leghasznosabbat/legkönnyebben kezelhetőt, 4-gyel a legkevésbé hasznosat/a legkevésbé könnyen kezelhetőt.

Hasznosság

feladatmegoldás _____
szókincs/frazeológia _____
szövegkezelési technika _____
nyelvhelyesség/helyesírás _____

Könnyen kezelhetőség

feladatmegoldás _____
szókincs/frazeológia _____
szövegkezelési technika _____
nyelvhelyesség/helyesírás _____

Appendix H Sample from the qualitative analyses

Excerpts from the think aloud data

Text: Think aloud data\Rater3
Position: 3 - 3
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Elég könnyű lesz megírni jól ezt a levelet. Na mindegy, majd odafigyelünk arra, hogy ne legyen benne átemelés.

Text: Think aloud data\Rater3
Position: 4 - 4
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Bár ezt át kellett fogalmazni, tehát nem egyszerűen csak átemelte.

Text: Think aloud data\Rater3
Position: 8 - 8
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Most jön ez, hogy minta útitervet kell adni, és javasolni ezt-azt. hát ugye az első két mondat az totál ugyanaz volt, mint a megadott szempontok. Eddig még semmit nem csinált.

Text: Think aloud data\Rater4
Position: 8 - 8
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Jajaj..hát…ez eddig lemásolta, amit a feladatban adtunk

Text: Think aloud data\Rater4
Position: 13 - 13
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
aztán ez megint másolás, de a kiegészítés rossz

Text: Think aloud data\Rater4
Position: 13 - 13
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
úgyhogy na hát most leírja, hogy mi lenne nálunk a feladata az új munkakörben, azaz lemásolja több-kevesebb sikerrel, azt, amit mi írtunk, ami a feladatban van

Text: Think aloud data\Rater4
Position: 13 - 13
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
igen, ezt sikeresen lemásoltuk, ide is ír …….. jó

Text: Think aloud data\Rater4
Position: 16 - 16
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Szókincs, frazeológia: nohát igen, hát amennyit önmagától használ, ha egyáltalán…

Text: Think aloud data\Rater5
Position: 3 - 3
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Hát a címzésbe nem tudunk belekötni, mert ezt ugye lemásolta a feladatról.

Text: Think aloud data\Rater5
Position: 6 - 6
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
a többit szépen kimásolta

Text: Think aloud data\Rater6
Position: 4 - 4
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting

hát így a közepe felé úgy látom, hogy nagyon erősen követi a meg…a javasolt pontokat.

Text: Think aloud data\Rater6
Position: 18 - 18
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
hát persze sok szó promptból jön, ugye az a….a mit tudom én a……meg végül is elég sok szó persze onnan jön…….ő

Text: Think aloud data\Rater7
Position: 6 - 6
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Ami nagyon nagy hiba viszont ezt követően, hogy ő… megismétli azokat a gondolatokat, itt fel… amelyek az utasításban benne vannak. Nevezetesen arra gondolok, hogy a pályázónak milyen feladatai vannak. Sőt továbbmegyek, nem meg… nem egyszerűen megismétli, hanem ugyanolyan ő… ???, tehát ugyanolyan ő… nem mondatokba öntött formába ismétli meg, hogy milyen feladatai lesznek majd, ha ne adj Isten megkapja ezt a munkát. Ahogyan ez itt fel van sorolva a feladatban. Ő… Hát ez két dologban is hibás, egyrészt azért mert ő… neki nem kell felsorolnia a feladatokat. Ez a kiindulás. Tehát neki ezt nem kell felsorolnia. Azt kéne bizonyítania, hogy ezeknek a feladatoknak az ellátására ő alkalmas. Ő… az önéletrajzból ugyanis kiderül, hogy neki milyen diplomája van, milyen ő… szakmai múlttal rendelkezik, de neki azt is bizonyítania kéne, hogy ő erre emberileg, szakmailag alkalmas. Tehát ő… ő… eleve nem a feladatokat kéne felsorolnia, hanem azt kéne bizonyítania, hogy alkalmas rá, továbbá az a legkevesebb, hogy ezt mondatokba kéne öntenie.

Text: Think aloud data\Rater7
Position: 6 - 6
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Ő… hát itt ő… ugye értelemszerűen nincsenek hibái, mert lemásolta az egészet úgy, ahogy van.

Text: Think aloud data\Rater10
Position: 2 - 2
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Ugyan kimásolt egy kicsit, de nem baj.

Text: Think aloud data\Rater10
Position: 13 - 13
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Na, azt mondja, hogy ….hol is…jó….vizsgázó a feladatnak megjelölt szövegfajtát…megvannak a dolgot….ezek elég szépek….ez megvan? …….hol van ez?.....igen, hát ezeket mind kimásolják, az a bosszantó,

Text: Think aloud data\Rater11
Position: 4 - 4
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
.ezt nyilván átemelte a feladatból

Text: Think aloud data\Rater13
Position: 11 - 12
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
Ezt köszönjük, hogy leírtad, csak ez abszolút kimásolt dolog innen a feladatmegadásból. Úgyhogy ez teljesen fölösleges itt…jó… ezt meg is hullámosozom…itt mellette…aztán…

Text: Think aloud data\Rater15
Position: 2 - 2
Code: 1.2 Criteria\1.4 Performance/text\1.4.4 Lifting
a legfeltűnőbb az volt, hogy a nagy részét kimásolta a feladatból

Text: **Think aloud data\Rater15**

Position: **3 - 3**
Code: **1.2 Criteria\1.4 Performance/text\1.4.4 Lifting**
ezt is kimásolta a Gabriella Gergely Baranyai.

Text: **Think aloud data\Rater15**
Position: **4 - 4**
Code: **1.2 Criteria\1.4 Performance/text\1.4.4 Lifting**
Igen, ez az a rész, amit egy az egyben kimásolt a feladatlapról…

Text: **Think aloud data\Rater15**
Position: **16 - 16**
Code: **1.2 Criteria\1.4 Performance/text\1.4.4 Lifting**
.címzésnél kimásolta kis után

Text: **Think aloud data\Rater15**
Position: **16 - 16**
Code: **1.2 Criteria\1.4 Performance/text\1.4.4 Lifting**
ő feladót úgy…szórul szóra úgy írta,

Excerpts from the interview data

Text: Interviews\I_1
Position: 55 - 55
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát ha üres volt a lap.

Text: Interviews\I_1
Position: 57 - 58
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Tintafelhasználás történt… ő… előfordult, igen. Hát előfordulhat, hogy leír három sort és és és… lehet, hogy előfordult. Szóval hogy jóváhagytam olyat.

Text: Interviews\I_2
Position: 97 - 98
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Nulla pontot nagyon-nagyon ritkán adok. Ha nem ír semmi mondjuk… ő… ha nincs jelentősége az 1 pontnak… semmilyen jelentősége nincsen az 1 pontnak abból a szempontból, hogy átmegy-e, vagy nem megy át. És ő… hát egyrészt a ti dolgotokat megkönnyíteni, hogy mit panaszkodjon, hogy… hát ez is hozzátartozik, ez a ……, kész.

Text: Interviews\I_2
Position: 101 - 102
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Tehát, ha ha ha megérdemelne valamit, vagy ilyen szempontból lenne tétje, akkor nem… tehát nulla pontot amiatt… és akkor ugye az egészet bukja. Azt semmi, ha nem írt semmit akkor mondjuk elfogadja és nem borítja rátok az asztalt.

Text: Interviews\I_2
Position: 105 - 106
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Valahogy egy pontot biztos, hogy adok. Arra, hogy írt valamit, leírt szavakat, és akkor ő azzal boldog. És és ő… meg szoktam gondolni azt is, hogy két pont, vagy három pont, mert neki mond valamit. Legközelebb próbálkozik és akkor… ha legközelebb próbálkozik és tanult és rosszabbat kap, akkor nekimegy a Dunának. Tehát igenis, hogy mondjam… nem úgy, hogy 5 pont alatt mindegy.

Text: Interviews\I_2
Position: 107 - 108
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Mert őneki… ott egy ember, előttem vannak mondjuk állandóan lelki szemeim előtt a kedves, gyenge diákjaim, akik igyekeznek, vagy… szóval a mögött is egy ember van, és ő… hát tudom, hogy neki fog számítani, ha legközelebb vizsgázik, hogy akkor elsőre kaptam 6 pontot és most tessék 4-et. Hát nem, akkor kapjon másodjára hetet. Tehát ő… nullát ritkán adok és ott font… szóval visszajelzés.

Text: Interviews\I_2
Position: 109 - 110
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Visszajelzés, egy embernek visszajelzés. Igen. Igen.

Text: Interviews\I_2
Position: 141 - 142
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát nullát, ha üres a papír.

Text: Interviews\I_3
Position: 9 - 9
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

Őszintén szólva nem hiszem, hogy adtam valaha 0 pontot, vagyis úgy hogy minden részpontszám 0, de szerintem olyat sem sokat, hogy csak egy részpontszám 0. Én nagyon harcolok azért, hogy csak az üres papír legyen a 0 pont. És megint nem csak azért, hogy a felesleges konfliktusokat elkerüljük. Tulajdonképpen úgy gondolom, hogy aki 0 pontot érdemel, az úgysem fog átmenni a vizsgán, és nincs nagy különbség a 0 vagy az 1 vagy a 2 pont között a végeredményt illetően. A vizsgázóra gyakorolt hatásban viszont igen. Egyszer egy kolléganőm mondott egy nagyon okos dolgot, és azóta is ez lebeg a szemem előtt, hogy a vizsgázót meg lehet buktatni, de megalázni nem kell. Szóval szerintem az 1 ponttal is meg fog bukni, de legalább nem alázom meg.

Text: Interviews\I_4
Position: 25 - 26
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

Most azt szeretném megkérdezni tőled, hogy ha ugye azt mondod, hogy egész mást ír, akkor attól kezdve nem nézed, mondjuk nem mondtak azt, hogy totálisan 0 pontot adsz, de akkor számodra van különbség a között, hogy egy üres papír, vagy egy olyan levél vagy egy olyan feladat, ami más?
Hát elvileg nincs. Elvileg nincs. Ő kapott egy konkrét feladatot, ami hozzá tartozott ahhoz… azokhoz az ismeretekhez. Tehát én szívem szerint tovább nem foglalkoznék vele. Nem ez a gyakorlat. Nem ez a gyakorlat, de ez bennem mindig is kérdés volt, és mindig is vitapont. Hogy hát akkor ne haragudj, megtanulom neked a Micimackót angolul és akkor tök mindegy mit kérdezel, megkapok rá egy felsőfokú nyelvvizsgát.

Text: Interviews\I_5
Position: 87 - 88
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

A… nulla ponttal… hát ugye ott vigyázunk, azért mert tudjuk, hogy a nulla pont, akkor… akkor bukta a vizsgát.

Text: Interviews\I_5
Position: 89 - 90
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

Ő… hogyha látom, hogy azért ő… azért írt csak másfél mondatot, hát nem látom, tehát sejtem, hogy azért írt, mert nem volt ideje átmásolni, némelyik oda is írja, hogy nem volt ideje, akkor adunk neki egy pontot, még akkor is, hogyha az gyakorlatilag értékelhetetlen lenne.

Text: Interviews\I_5
Position: 99 - 100
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

Kivéve, ha nullát. Igen, ez igaz. De már az is mindegy szerintem. Tehát hogyha a… a levél azért rengeteget elárul. Tehát aki nem bír megírni egy épkézláb levelet, valószínű a tesztnél is a szókincs… vagy a nyelvtani résznél a… a 25 itemből talán ötöt-hatot eltalált véletlenül.

Text: Interviews\I_5
Position: 101 - 101
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero

A szóbelin meg biztos el fog vérezni, tehát ő… így, hogy több részből áll a vizsga, az emberbe benne van az, hogy na legyen, adok három pontot, mert úgyse megy át.

Text: Interviews\I_6
Position: 23 - 23
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
A nulla ponthoz azt kell tenni, hogy semmit se szabad írni. Vagy mondjuk éppen ott a megszólítás, mondjuk

Text: Interviews\I_6
Position: 25 - 25
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Van különbség egy üres lap, meg egy kis irka-firka között, bár előfordulhat, hogy az irka-firka is nulla pont.

Text: Interviews\I_6
Position: 27 - 27
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Előfordulhat. De… de hogyha… hogyha már van kerete, ugye tehát megszólítása, dátuma, mit tudom én, szóval… akkor már ugye kap egy pontot legalább. Vagy… vagy ha valamennyire kiderül, hogy tulajdonképpen milyen irányban akar közölni valamit, akkor azért egy, egy-egy… lehet, hogy csak egy pontot, vagy két pontot, de azért azért tulajdonképpen nullát, azt azt vagy üres a lapja, vagy ő… vagy tényleg három olyan sor van rajta, amiből… ami hottentottából is lehetne.

Text: Interviews\I_7
Position: 19 - 19
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát ha nulla pontot ér… végül is nagyon nehéz adni az értékelési skála szerint. Mert már önmagában a feladatmegold… egyedül talán a feladatmegoldásnál lehet, hogyha véletlenül félreértette a feladatot és mondjuk ajánlatkérés helyett ő… ajánlatot írt. De az összes többinél borzasztóan nehéz nulla pontot adni, mert szókincse akkor is lesz és hogyha nem ajánlatkérést írt például, hanem ajánlatot, akkor is nagyjából ugyanaz a szókincs fog előfordulni, tehát nem mondhatom azt, hogy nem használta a szakszókincset, még ha egy picit más is a szókincs, mert mondjuk ajánlatkérést reklamációval biztos nem fog összekeverni. Nyelvtant fog használni, tehát kénytelen vagyok valamilyen pontot adni rá. És hát még a szociokultúrális kompetencia.

Text: Interviews\I_8
Position: 33 - 33
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
A… a bukásnál hát nyilván, hogyha nem írt semmit, akkor az egyértelmű dolog, hogy az nulla, vagy csak két szót írt oda, akkor az nyilvánvalóan nulla. De vannak olyan … olyan nagyon… nagyon rosszul sikerült írásbelik, például ez a szakmai… szakmai kérdésekre adandó kis válasz, aminek se füle, se farka, vagy lehet, hogy… hogy oda van írva egy értelmes bekezdés, de egyáltalán nem arra válaszol, amit kérdeztünk. És ő… szóval az ilyen … ilyen rossz megoldások azok… azok azért erősen arra ingerelnek, hogy nullát adjak.

Text: Interviews\I_9
Position: 37 - 37
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát ő... hm… valószínű… valószínű, hogy a nullába nem fér bele az, hogy ő… hogy egészen nulla legyen. Ha valami egészen másról ír értelmese, mert valamit… valamit felmutatott. Igen. De mondjuk azt, hogy a nagyon gyenge megoldások azok azok éppúgy épp úgy nullásak a szememben, mintha semmit nem írt volna oda.

Text: Interviews\I_10
Position: 39 - 39

Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát a nulla azt hiszem, hogy nem, mert ugye a tendencia, amit úgy egymástól átveszünk az talán az, hogy ő… hogy csak akkor adunk nullát, ha üres a papír. Ő… mert ha már odarakott egy… megszólítást, vagy nem tudom mit, akkor… hát nem tudom. Nyilván erre még nullát kéne adni, de… de a… már valamit írt, szóval a nulla az nagyon ritkán.

Text: Interviews\I_12
Position: 33 - 33
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Előfordult sokszor, hogy nulla pontot adtam amikor azért produkált valamit. És ugye az az elv, hogy ha valamit már leírt, vagy valami… kosz van a papíron, akkor erre ne adjunk nulla pontot. Hát most ebbe nem érdemes belemenni, hogy melyik az a határ, ahol, ami fölött már lehet nem nulla pontot adni. Ő… sokkal, szóval akkor érzem a nulla pontot, amikor még ha produkált is valamit, de az valami olyan elképesztően semmit mondó és és ő hiába írt le valamit és produkált valamit, annak az ég világon semmiféle ő… kommunikatív hát értéke nincsen, akkor ő én szeretek nulla pontot adni.

Text: Interviews\I_12
Position: 47 - 48
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Hát meg nem esküszöm rá persze, de ha nyilván, hogy ha mindegyikre nullát… az az üres papír. Nem, azt hiszem nem fordult még elő.

Text: Interviews\I_13
Position: 19 - 20
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Kezdjük a nullával. Tehát mit kell a nulla pontért tenni?
Alapvetően az, hogy nem ír semmit.

Text: Interviews\I_13
Position: 22 - 22
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Az a legelső, hogy nulla pont. Illetve… hát feladatmegoldásra nem tudok nulla pontot adni, csak… csak akkor, ha nem írt semmit. Ha jól emlékszem. Tehát ott mindenképpen azért egy pontot kellene adni, bár most nincs előttem teljesen a skála. És… azért nulla pont, hogyha… ha abszolút értelmezhetetlen, ha nyelvtanilag is nagyon… nagyon ő…abszolút a határ alatt van és ha nagyon kevés. Tehát ha összességében nézem a produkciót és az a benyomásom róla, hogy ez… ez így aztán értékelhetetlen és elfogadhatatlan, nyilvánvalóan, hogyha írt valamit, akkor már nem tudsz neki nulla pontot adni, mert hogy még akkor is, hogyha másról írt, akkor is valamit értékelsz a szókincsben, a szövegkezelésben vagy a nyelvhelyességben. De hogyha együtt van ez a három, hogy ő… se nyelvtanilag és nyelvhelyességileg, se szövegkezelésben, mert mondjuk írt három mondatot és azt is stílusában úgy, hogy az abszolút nem felel meg, és hát szókincs meg sehol nincs, akkor… akkor muszáj neki rossz pontokat, adott esetben nullát kapnia.

Text: Interviews\I_13
Position: 24 - 24
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Dehogynem. Hát az üres papír az egyértelműen végig nulla pont. Mind a négy… négy pontnál.

Text: Interviews\I_13
Position: 25 - 30
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Ezt értem, de némi kis irka-firkával is lehet nálad nulla pontot elérni?
Igen. De… tehát nem mind a négy szempontban.
Ühüm.
Hanem a feladatmegoldásban mondjuk és a… mit tudom én most mondom az, hogy a szövegkezelésben.
Ühüm.
De a másik kettőre lehet, hogy kap pontot.

Text: Interviews\I_14
Position: 25 - 27
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Ritkán. Hát természetesen amikor nem írt semmit…
Na jó, üres a papír. Ezen kívül.
Üres a papír, vagy a papíron tényleg csak egy név van, vagy a címet írta le, tehát az is gyakorlatilag üres papír. Ő… hát a nulla ponttal ő… azért szoktam enyhébben ő… eljárni, hogyha már mondjuk egy bevezető mondat, vagy egy paragrafus ő… megvan, mert ő… mert tudom, hogyha nulla pontja van, akkor az egész vizsgája nulla lesz. És most a felülbírálatnál is pont van ilyen, hogy 98 pontos ő… anélkül, hogy az egyik feladat, ami nulla pontos, anélkül is elérte a 98 pontot. És ő… és ezek mindig bántják az embert, hogy vagy nem volt ideje, vagy még az is előfordulhat, hogy esetleg nem arra a lapra írta, hanem magára a feladatra, tehát itt nagyon sok variáció és figyelmetlenség is benne lehet.

Text: Interviews\I_14
Position: 87 - 87
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
és például a… a nulla pontosak. Tehát az elég nehéz eldönteni a nulla pontot a sávleírások alapján. Mi az, hogy szóhasználata érthetetlen, vagy másról ír, vagy nem ír semmit? Tehát ő… ezek nagyon sokszor… nem mondhatja az ember egyértelműen, hogy nem ír semmit, mert tele van a papír, az sem írhatja, hogy másról ír, mert valamennyire betartotta a feladatot, ő… de ő… de de mégsem egy pont.

Text: Interviews\I_14
Position: 91 - 91
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Tehát valahogy a nulla sávnak a leírását ő.. szóval nehéz ő… ilyen mondatai… mondatai értelmezhetetlenek. Hogy annak milyen szintjén értelmezhetetlen az egyeshez képest? Igen, tehát nekem itt alul elég nehéz, az alsó sávok..

Text: Interviews\I_15
Position: 20 - 20
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
lehet, hogy nem lesz népszerű amit mondok, de… de ő… nulla pontot ő… nem szoktunk adni, ő… azt lehet mondani, hogy… hogy soha. Nullát akkor adunk, ő… ez a mi gyakorlatunk ő… a német javításnál, nullát akkor adunk, hogyha a papír teljesen üres.

Text: Interviews\I_15

Position: 26 - 26
Code: 2 Extremism, central tendency\2.2 Reasons for zero/maximum\2.2.1 Zero
Szóval azért ez nem ennyire így működik. De hát miért írna le olyat, ami semmi köze

nincs, tehát ő… és írásfeladatnál ilyen még soha nem fordult elő, hogy ő… hogy ő… az alsó, szélsőséges értéket kellett volna adni. Az előfordult, hogy feladatmegoldásra nullát kellett neki adni, mert teljesen mást írt. És akkor ilyenkor mindig bajban vagyunk, erről már máskor is beszéltünk, hogy… hogyha egyszer a feladatmegoldás nulla, akkor milyen alapon adunk mi másra pontokat? De hát azért mégis írt ő… akár egy 150 szavas fogalmazást is, csak tökéletesen nem arról írt. És akkor… akkor mindig, ezen mindig elvitatkozgatunk, hogy akkor most a többi feladatnál mennyit vonunk le, hogy vonunk le? Ő… de például a nyelvhelyességnél, ha netán az illető ő… négyes nyelvhelyességet írt, akkor ott még csak le se lehet vonni.