Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar

An Investigation of Rater and Rating Scale Interaction in the Validation of the Assessment of Writing Performance

A javító és az értékelőskála interakciójának vizsgálata az írásfeladatok értékelésének validálása során

című PhD doktori értekezéséhez

Dissertation summary

A doktori (PhD) értekezés tézisei

Benke Eszter

Témavezető: Dr. habil. Kormos Judit

Neveléstudományi Doktori Iskola

Nyelvpedagógia program

Budapest, 2007

Contents
Tartalomjegyzék

1 Theme and aims of the dissertation

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 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.

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?

2 Overview

Chapter 1 introduces the broad context of the research which is further elaborated in the subsequent chapters. The theoretical and practical aims are outlined along with the rationale underlying the investigation. This chapter also lists the two major and the six 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 are discussed. The second part of the chapter focuses on measurement error: from the most general theoretical approach to measurement error, the discussion moves on to its specific form prevalent in the assessment of subjectively scored tasks. This is 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 are 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 are reviewed. Although my focus is rater and rating scale interaction in subjective assessment, other aspects of the rating process might also have an indirect impact on this interaction. The brief overview of validation methods serves 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 concludes 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 methods of data analysis are explained in more detail. As only a limited amount of the possible output that such analyses can yield is used and interpreted, the description of the IRT based data analysis is also 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 are also highlighted besides listing further unmapped areas worth investigating in relation to the topic.

3 Theoretical background

As the aim of the study was to establish the validity of an assessment instrument by identifying possible sources of measurement error in the assessment of writing, there were three general theories the paper built from in order to provide a systematic approach to the research questions. The first is measurement error as conceptualised in classical true score theory (Crocker & Algina, 1986; Spearman, 1904) and modern test theory (Carmines & Zeller 1979; Crocker & Algina, 1986; Nunnally & Bernstein, 1994; Viswanathan, 2005).

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 (Linacre, 1989), which permits a relatively objective approach to subjective assessment. Thus, in an attempt to establish the validity of the assessment procedure, the second major theoretical model informing the research is Many-faceted Rasch measurement (Linacre, 1989). This theory is an extension of the one-parameter Rasch model which is capable of modelling facets of interest other than task difficulty and examiner ability. The model 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. Rater error (Engelhard, 1994; Linacre, 2003-6; Saal, Downey & Lahey, 1980; Wolfe, Moulder, Bradley & Myford , 2001) is the special focus of the current study as rater variability and diverging rater characteristics raise concerns regarding the validity and the reliability of the measurement procedure in rater-mediated assessment.

Thirdly, as the focus of investigation is the assessment of writing, influential models of writing assessment (Engelhard, 1992; Lumley, 2005) also informed the line of investigation. These models identify the intervening variables which, during the measurement process, provide a link between measured proficiency and observed rating. These performance dimensions highlight sources of variability and measurement error in performance assessment which can be identified and controlled for with the application of MFR analysis.

4 Research methods

To answer the research question, the study addressed the issues proposed above in two stages, following both quantitative (Study 1) and qualitative (Study 2) methodology.

4.1 Data collection and analysis in Study 1

Altogether scores awarded on 2011 scripts by 27 raters on seven occasions during a three-year period constituted the data for the analysis in the first study. Many-faceted Rasch analysis was carried out with the help of FACETS (Version 3.61.0) software. For the analysis applied in this quantitative part of the study, the subscores awarded on each criterion were used. Instead of the final agreed scores, the individually awarded scores were analysed as these provide a more authentic and accurate reflection of the use of the rating scale.

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. Does the factor structure of the total scores confirm the appropriate functioning of the 6-point analytic 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.

4.2 Data collection and analysis in 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 raters' perceived rating behaviour. 15 raters took part in this part of the study, nine of them were teachers of English, and four were teachers of German. 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.

The second study, thus, with information obtained from the verbal protocols and the interviews sought to provide rich data for the following set of 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?

Table 1 provides a brief summary of the methods of data collection and analysis for each research question.

Table 1 Research methodology matrix

Research question	Source of data	Method of data analysis
Which assessment criteria generate bias of rater behaviour? Which criteria elicit little variation in the distribution of the awarded scores? To what extent is the halo effect or the cross-contamination of descriptor bands apparent in the distribution of scores? Does the factor structure of the total scores confirm the appropriate functioning of the 6-point analytic rating scale?	Scores of 2011 scripts awarded by 27 raters on an analytic rating scale Data collected on 7 rating occasions during a three-year period	Many-facet Rasch analysis to identify rater, category and criterion misfit
Why do assessors exhibit different rating profiles across different domains of the rating scale? What construct-irrelevant factors emerge during the application of the rating scale?	Concurrent verbal protocol carried out with fifteen raters during the rating process of three sample scripts; Interview conducted with fifteen raters on perceived rater behaviour	Qualitative data analysis with MaxQDA to identify rater misbehaviour

5 Major findings and their implication

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

5.1 Evidence for rating scale validity

The first broad research question sought to provide 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). 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, significant biases should be dealt with regardless of whether they advantage or disadvantage the candidate.

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. 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. The high reliabilities of the separation indices also confirmed that raters are capable of differentiating between the rating categories.

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.

5.2 Insight into rater behaviour

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.

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

2. What construct-irrelevant factors emerge during the application of the rating scale?
It seems from the interviews that there are more factors which are conducive to generous rater attitude than factors that trigger a negative approach. Although no deviation from 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. Raters' generosity

might increase the number of false positives, whereas harshness would contribute to the emergence of false negatives. Neither of them is 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.

5.3 Practical benefits of the study

The results raised 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, 2004). 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.

6 Conclusion

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.

7 Contribution of the dissertation

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. In addition, as the rich data obtained from Study 2 suggest, a more extensive use of qualitative validation methods can give a deep insight into rater behaviour, and latent sources of perceived or actual rater misbehaviour can be revealed. The concept of rater validity and the constant monitoring of rater behaviour can also significantly contribute to the improvement of the evaluation procedure and the elimination of measurement error. These are the areas where the present study hopes to add something new to the body of existing research.

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.