E8 Writing

Validation Report for E8 Writing

LTC Technical Report 6

Jutta Götzinger

Bundesinstitut bifie

CONTENTS

1. Introduction

The rating of writing performances is time consuming and cost-intensive work as it involves essay scoring by human raters. In order to reduce costs and time computers could be used in place of or in addition to human raters. The use of computers in scoring was first introduced by Page in the 1960s and has been further developed up to the present. The question that can be raised is whether automatic scoring programmes tend to rate performances similarly to the human raters. Schwartz (in Shaw and Weir 2008: 212) states that 'apart from being cost effective, computerised scoring is unfailingly consistent, highly objective and almost wholly impartial'.

The aim of this study is to investigate textual features in the short and long E8 performances that predict the ratings of human raters and which show that automated scoring procedures are also scoring the underlying writing construct. In particular this study focuses on the investigation of predictor variables for the following E8 rating dimensions: Coherence and Cohesion, Grammar, Vocabulary.

2. Why is validity a central concept in language testing?

The main purpose of language testing concerns the measurement of test takers' language abilities and hence the interpretation of test scores and the inferences we make on the basis of these test scores. According to Bachman, scores on a test can be used as indicators of a test taker's language abilities as long as these scores are both reliable and valid (Bachman 1991: 24). A crucial prerequisite for a 'good' test is *reliability*. Bachman and Palmer define reliability as 'consistency of measurement' (1996: 19). Therefore, a test is said to be reliable if its scores are consistent among different features of the testing situation (Bachman, Palmer 1996: 19).

The second even more crucial quality in language testing addresses the issue of *validity*. Henning defines the concept of validity in the following way:

> "Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term valid when used to describe a test should usually be accompanied by the preposition for. Any test then may be valid for some purposes, but not for others." (1987: 89)

Thus, the main question concerning the investigation of validity in language testing is: "Does the test accurately measure what it is intended to measure"? As Bachman argues, we have to bear in mind that 'the inferences or decisions we make on the basis of the test scores are meaningful, appropriate and useful' (1991: 25). In order to ensure validity of a test we should take into account the purpose of the test. For example, if we administered a writing test compiled of multiple choice questions, in order to make inferences about test takers' writing abilities, the test scores would not be valid, as test takers do not produce a sample of writing performance which represents their writing abilities, for instance paragraphing or linguistic and textual knowledge. Hence, for the sake of validity, it is essential to clearly specify what we want to measure in a particular test. This means that if test scores are used as meaningful indicators of one's language ability, we must be sure that these scores are not affected by factors other than those which we intend to measure (Bachman 1991: 25).

3. Aspects of validity

Traditionally the concept of validity has been divided into different types and it can be investigated in various ways. Alderson et al. emphasise that the different types of validity should be seen as different methods of gathering evidence of a test's validity (1996: 171). The most common types of validity are outlined below.

3.1. Face validity

Face validity may be considered as the weakest type of validity. A test that is said to have face validity 'looks as if' it measures what it is intended to measure. As Alderson et al. state, 'face validity involves an intuitive judgement about the test's content by people whose judgement is not necessarily *experts*' (1996: 172). Generally, these judgements are holistic, but the focus may also be on single items or ambiguous instructions in order to obtain a global judgement about a particular test. For example, a test that purports to measure writing ability but which does not require the individuals to write might show a lack of face validity. The concept of face validity is seen rather critically by many professional language testing researchers and

regarding the usefulness of this approach opinion is divided amongst experts (Sigott, 2004: 45).

3.2. Concurrent validity

In contrast to face and content validity, *concurrent validity* is concerned with the relationship of test scores obtained on a newly developed test and another existing criterion measure. The existing criterion may be a well-established standardized test, which is considered to be valid, or a set of judgements which were made after the test was administered. If there is a high correlation between the criterion measure and the test scores, the newly developed test is said to measure the same type of language ability as the existing criterion and therefore the test is considered to be valid. We can distinguish between *concurrent* and *predictive* validity, depending on whether the scores on the existing criterion measure are gathered at the same time with the new test scores or subsequently (Sigott 2004: 46).

However, Sigott argues that 'demonstrating that a test ranks the same subjects in the same way as another test does not mean specifying the abilities which are required to perform well on either test' (2004: 46). It rather indicates that the skills for performing well on both tests are similar or even identical. Furthermore, this approach does not specify whether the correlation between the two tests is due to test content or results from test method effects (Sigott, 2004: 46f).

3.3. Content validity

The next form of evidence refers to the content of a particular test. The concept of *content validity* is clearly distinguishable from face validity as the former uses a theoretical basis for examining test content. As Davies states, the approach of content validity is a conceptual or non-statistical one that is based on a 'systematic analysis of the test's content investigating whether it constitutes an adequate sample of the language skills or structures to be measured (1999: 34). These structures and language abilities may be clearly stated in test specifications or can be based on a syllabus. Both definitions of test content may be used as a basis for the validation process. Hughes (2007: 26) points out that it is not necessarily the case that everything described in the specifications will always be demonstrated in one single test. In order to judge whether a test has content validity Hughes suggests comparing the test specifications and the test content and adds that these judgements should

preferably be made by people familiar with test construction (2007: 27). However, these people should not be directly involved in the construction of the test in question (2007: 27).

3.4. Construct validity

Among all types of evidence, *construct validity* is regarded as the most complex type of validity since it includes both analysis of the test from the content point of view and an analysis of test performance and test scores (Sigott 2004: 47). The idea of construct validity refers to the degree to which 'performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs' (Bachman 1991: 255). In examining construct validity the first question to be addressed is 'What are the abilities that we want to measure?' When we define what we want to measure, i.e. specified abilities according to the test's purpose, we are defining a construct. As these theoretical abilities are not observable directly, a test is used for their operationalisation. The difference between construct validity and content validity is that the former uses a theory 'to explain or predict subjects' behaviour on the test rather than to serve as a basis for ensuring adequate coverage or sampling of the test content' (Sigott 2004: 47).

Evidence for a test's construct validity can be obtained in different ways. According to Alderson et al., one way to evaluate construct validity is to examine the relationship of different test components with each other (1996: 184). We assume that each of these test components measures different language abilities. Consequently the correlation coefficient should be fairly low, as a high coefficient would indicate that the different test components measure the same kind of ability.

Another approach to construct validation is to compare test performance with psychological data or biodata, obtained from students at the time of test administration (Alderson et al., 1996: 185). The purpose of this approach is to find out whether certain psychological characteristics or biodata characteristics affect students' performances.

According to Bachman (1991), *multitrait-multimethod analysis (MTMM)* is a classical approach to construct validation. The underlying concept of this approach is a combination of internal and external validation procedures. Tests which show a significant relationship among each other will demonstrate higher intercorrelations

(convergent validity) than tests that do not show any relationship (divergent validity)(Alderson et al., 1996: 186). Convergent validity refers to the degree to which different measures of the same trait are inclined to agree whereas divergent validity shows the extent to which measures of different traits tend to produce different results (Bachman 1991: 263).

To sum up, evidence for a test's validity can be gathered by applying different validation procedures: face validation, concurrent validation, content validation and construct validation. Sigott 2004 points out that 'none of these aspects and types of evidence is sufficient in itself' (2004: 43). Further, Alderson et al. emphasise that 'the more different 'types' of validity that can be established, the better, and the more evidence that can be gathered for any one 'type' of validity, the better'. (1996: 171). In other words, a test should be validated in as many ways as possible.

In the case of the E8 Writing Test construct validation is carried out. One approach to construct validation in this study is the investigation of how well linguistic features elicited from test performances accord with the test construct. In the context of the E8 Writing Test, an important aspect of construct validity has to do with whether scores of various linguistic characteristics (produced by a quantitative automated analysis) truly measure the test takers' writing ability. The validity is based on how closely quantitative analysis produced by a computer program can predict the scores given by human raters. Though writing in the E8 context is judged from four perspectives, task achievement, coherence and cohesion, grammar and vocabulary, the current study investigates selected aspects of grammar and vocabulary only.

4. Method

As mentioned above, the primary intention of a language test is to make inferences about test takers' language abilities. A crucial factor therefore is to clearly specify *what* the test is intended to measure and how this *what* is implemented in a language test. In other words we need to specify the construct of the test in order to know which abilities are measured.

The aim of the E8 Writing Test is to diagnose strengths and weaknesses in test takers' writing competence so that teachers can adapt their instructions to meet their pupils' needs. The E8 competence model includes four main parts which are all of

equal importance. The primary competence measured in this test is the communicative competence which is shown in an appropriate response to the task (Gassner, Mewald, Sigott 2008: 20). Furthermore it is the competence to produce fluent text which demonstrates coherence and cohesion on both, sentence level and paragraph level. Thirdly, a good range of grammatical structures and the competence to use them accurately is measured. Fourthly, a good range of vocabulary relevant to general topics and the accurate use of words are tested.

The writing test itself is divided into two sections. The first section comprises a short writing task with an expected response of 40 to 60 words whereas in the second section, the long task, test takers are asked to write 120 to 150 words. The total testing time is 45 minutes, 10 minutes for the short task, 20 minutes for the long task, 5 minutes for revision and 10 minutes for administration at the beginning and at the end. The instructions for this test are written in simple English and they are sufficiently detailed so that test takers exactly know what is expected of them. Weigle (2002: 103) points out that clear instructions which state the specification of the audience and the purpose of writing are an important aspect of a valid test. The E8 Writing Test instructions also include the required length of the texts indicated in words.

The prompts of the E8 Writing Test are developed by practising teachers of English. Before being administered, all prompts have to undergo a screening process which ensures that the prompts are adequate for the test takers' age and their language level. This is no higher than A2 in the CEFR since the difficulty level of the test is supposed to encompass levels A2 to B1 in the CEFR. Furthermore the prompts are designed to be free of stereotypes since no test taker should be at a disadvantage. Some prompts may also contain pictures or drawings.

4.1. The subjects

The subjects are Austrian pupils who attend grade 8 in General Secondary School (Allgemeinbildende Pflichtschule (APS) or in Academic Secondary School (Allgemeinbildende Höhere Schule (AHS). All three ability groups in APS are tested. Most of the test takers have reached the age of 14 when the test is administered. Each test taker has submitted one short and one long performance. For the analysis

in this study a sample of 30 was randomly selected from the scripts produced during the 2007 administration. A total of 839 candidates took the E8 Writing Test in 2007.

4.2. The rating scale

In the case of the E8 Writing test, the existing analytic rating scale represents the aspects of writing that are considered to be part of the construct (Task achievement, Coherence and Cohesion, Grammar and Vocabulary). The descriptors of the E8 rating scale are graded using adjectives like 'good', 'generally sufficient', 'limited' or 'extremely limited'. All four rating dimensions consist of eight bands ranging from zero to seven. The last three dimensions are linked to the CEFR and the Austrian *Bildungsstandards Englisch* whereas there is no suitable category for task achievement in the CEFR. Weigle (2002: 120) points out that one of the advantages of using an analytic scoring schemes over a holistic schemes in a second-language environment is that it 'provides more useful diagnostic information about students' writing abilities'. Another advantage of using an analytic scale for rating second language performances is that analytic scoring can be more reliable than holistic scoring as 'reliability tends to increase when multiple scores are given to each script'. The test takers' writing performances are handwritten on two separate A4 sheets. Each script is marked by at least one assessor who went through a specific one-year rater training programme.

4.3. The raters

In addition to a valid rating scale the consistent application of rating scales is considered as a key parameter in the valid assessment of second language performances (Shaw & Weir 2007: 168). Elder (in Shaw & Weir 2007: 169) states that 'Subject specialists and language trained EFL teachers demonstrate a tendency to employ rating instruments differently'. In this study the scripts were rated by teachers of English who teach in lower secondary schools. The rater characteristics of these raters may differ in some ways, as some of these teachers are native speakers now living and working in Austria, or as they have different educational backgrounds. Some teachers have a university background whereas others attended a College of Teacher Education. In terms of rater training there are no differences

among the raters, as all of them had to do a one year training programme which started in October and prepared the assessors for the rating session in June. The aim of this rater training was to familiarise them with the rating scale and the rating process itself. In terms of harshness or leniency raters may interpret the rating dimensions differently. In order to adjust for these differences and differences in task difficulty multifaceted Rasch analysis was used.

4.4. The feedback

As the E8 Writing Test is a diagnostic test diagnostic feedback is reported on the four dimensions of the Writing Scale (Task Achievement, Coherence and Cohesion, Grammar, Vocabulary) ranging from 0 to 7 with reference to the CEFR up to level 'A2 above'. As already mentioned, differences in rater severity and task difficulty are adjusted for by means of multifaceted Rasch analysis. Therefore, the results are comparable across all pupils regardless of which assessor rated the performance and what prompt the script was based on.

4.5. Variables

As mentioned above, the main aim of this study is to find out whether selected text variables, representing aspects of the rating dimensions, can be used as predictors for the ratings of human raters.

To obtain measures for the variables a sample of 30 short and 30 long out of 839 E8 writing performances was chosen. All these 60 handwritten performances were Word-typed and obvious spelling mistakes were corrected as misspelt words would have had an influence on the quantification of some of the variables (i.e. type-token ratio, average word length based on tokens, average word length based on types). The variables were chosen with reference to the E8 Writing Rating Scale.

4.5.1. Coherence and Cohesion variable

The dimension of cohesion and coherence was measured by calculating a cohesion ratio. This means that the number of connective devices which are all mentioned in the E8 rating scale (i.e. *and, then, but, because*) were counted manually. The total number of connectors (tokens) was divided by the total number of words used in the text. No distinction was made between coherence and cohesion at paragraph level in contrast to sentence level. Furthermore, this study does not investigate the appropriate use of connective devices. It only focuses on the quantity of connectors. It is expected that more proficient writers use connective devices more frequently and may therefore get a higher rating on coherence and cohesion.

4.5.2. Grammar variable

In the E8 Writing Rating Scale grammatical competence ranges from 'a good range of structures for most communicative needs' on level 7 to 'an extremely limited range of simple structures or patterns within a learnt repertoire' on level 1. Average sentence length was chosen as a predictor variable for range of syntactical structure. It is expected that longer sentences are more complex in syntax and therefore result in higher ratings.

4.5.3. Vocabulary variables

In the case of the E8 Writing test the main interest lies on the test takers' range of vocabulary and the accurate use of the words. First, the *type-token ratio* was chosen as an indicator for the range of vocabulary used. The type-token ratio is the ratio of different words used in the text (i.e. types) and the total number of running words (i.e. tokens). A high type-token ratio means that a performance consists of a wide range of different words whereas a low figure shows that the writer has used few different words, which are frequently repeated (Read 2000: 203). Read points outs that

> it is reasonable to expect that more proficient writers have a larger vocabulary knowledge that allows them to avoid repetition by using synonyms, superordinates and other kinds of related words. (2000: 200)

In this study each inflected word form that occurs in the performance is counted as a different type. For example, *tell* and *tells* are counted as two different types.
One complicating factor regarding the type-token ratio (TTR) is the different length of the texts since the TTR-value is dependent on the length of a text. Therefore it is difficult to make texts of different length comparable. The WordSmith programme, developed by Mike Scott (University of Liverpool) is able to solve the problem of different text lengths as it counts every **n** words, then the ratio is calculated afresh and so on to the end of the text. The program computes a running average, which means that it yields an average type-token ratio which is based on n-words of consecutive word chunks. This ratio is called standardised type-token ratio (STTR). By calculating the standardised type-token ratio, it is possible to compare TTR-ratios across texts of different length (Scott, 2004).

The second variable is the *average word frequency on the basis of types* and *the average word frequency on the basis of tokens.* As the type-token ratio only comprises the quantity of different words used, the average frequency of a word was included in order investigate the rarity of the words. This variable indicates how many frequent or infrequent words are produced by the test takers. The basis for the analysis of students' word knowledge was the British National Corpus. This corpus includes written and spoken texts from various genres. For the frequency analysis the programme *BYU-BNC: British National Corpus* which was developed by Mark Davies was used.

The next variables that were chosen to function as predictors of the vocabulary rating are the *average word length in characters on the basis of types* and the *average word length in characters on the basis of tokens*. Frey and Heringer argue that the average word length measured in characters is a stable parameter for predicting human ratings of writing performances. (2007: 339)
In the current study all proper names and numbers were excluded from the analysis. Hyphenated words, for example "*I've"* were set to have three characters. First, the average word length on the basis of types was determined. For example, *and,* which consists of three characters, was only counted once as a word consisting of three characters. The second parameter, *average word length based on tokens*, also

included the number of occurrences of a certain word in the text, i.e. if "*and*" occurred twice in a performance it was seen to have six characters.

Hypotheses:

Based on previous research and literature the following hypotheses were formulated:

1. The cohesion ratio will predict the ratings of coherence and cohesion. There will be a relationship between the human ratings and the cohesion ratio in the short and in the long performances.

2. The average sentence length will predict the ratings of grammar in the short and in the long performances.

3. Standardized type token ratio, average word length in characters on the basis of types and on the basis of tokens, average word frequency in the BNC based on the types and on the tokens will predict the vocabulary ratings in the short and in the long performances.

5. Results

The main aim of this study is to find predictor variables for the rating dimensions of the E8 rating scale. In this section, the descriptive statistics of the predictor variables are presented. Three types of data analysis were introduced. Firstly, multi-faceted Rasch analysis which adjusted the raters' harshness and task difficulty was carried out before. Secondly, Pearson product-moment correlations coefficients were calculated in order to find out whether there is a relationship between two or more variables. In this study we want to investigate whether the human ratings as one variable tend to covary with another variable that emerge from text analyses.

The prerequisite for the third kind of analysis, regression analysis, is some relationship between the rating dimensions and the predictor variable. If the correlation analysis shows that one variable tends to have a relationship to another variable, we might be able to predict the future value of a variable.

5.1. An overview of correlation statistics.

5.1.1. Correlations among the rating dimensions

Table 1: Intercorrelations short performances

Correlations

		TA_Fair. Avge_short	COH_Fair. Avge_short	GR_Fair. Avge_short	VOC_Fair. Avge_short
TA_Fair.Avge_short	Pearson Correlation	1,000	,346	,350	,569**
	Sig. (2-tailed)		,061	,058	,001
	N	30	30	30	30
COH_Fair.Avge_short	Pearson Correlation	,346	1,000	,651**	,566**
	Sig. (2-tailed)	,061		,000	,001
	N	30	30	30	30
GR_Fair.Avge_short	Pearson Correlation	,350	,651**	1,000	,732**
	Sig. (2-tailed)	,058	,000		,000
	N	30	30	30	30
VOC_Fair.Avge_short	Pearson Correlation	,569**	,566**	,732**	1,000
	Sig. (2-tailed)	,001	,001	,000	
	N	30	30	30	30

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

From tables 1 and 2 it can be seen that the intercorrelations of the short performances are in general lower than those of the long performances. In contrast to the long performances, the short performances only show significant correlations among task achievement and vocabulary (0.569), coherence and cohesion and grammar (0.651) and vocabulary (0.566) and grammar and vocabulary (0.732). The

correlation analysis shows that there are no significant correlations among task achievement and coherence and grammar.

Table 2: Intercorrelations long performances

Correlations

		TA_Fair. Avge_long	COH_Fair. Avge_long	GR_Fair. Avge_long	VOC_Fair. Avge_long
TA_Fair.Avge_long	Pearson Correlation	1,000	,641**	,689**	,880**
	Sig. (2-tailed)		,000	,000	,000
	N	30	30	30	30
COH_Fair.Avge_long	Pearson Correlation	,641**	1,000	,661**	,846**
	Sig. (2-tailed)	,000		,000	,000
	N	30	30	30	30
GR_Fair.Avge_long	Pearson Correlation	,689**	,661**	1,000	,891**
	Sig. (2-tailed)	,000	,000		,000
	N	30	30	30	30
VOC_Fair.Avge_long	Pearson Correlation	,880**	,846**	,891**	1,000
	Sig. (2-tailed)	,000	,000	,000	
	N	30	30	30	30

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

As can be seen from table 2, the correlation coefficients range from 0.641 (Task achievement and Coherence and Cohesion) to 0.891 (Grammar and Vocabulary). For the long tasks the correlation among the four dimensions is highly significant. From this statistic output we might conclude that the E8 Writing Test measures one single trait, i.e. writing competence in terms of task achievement, coherence and cohesion, grammatical knowledge and vocabulary knowledge.
In the short performances as well as in the long performances it is noticeable that the highest correlation is between grammar and vocabulary.

5.1.2. Relationships among the rating dimensions and the predictor variables

After the relationship between the different rating dimensions was analysed, it was further of interest how the set of predictor variables performed in the correlation analyses. First of all, it could be observed that some predictor variables do not predict the rating dimensions as expected.

5.1.2.1. Short performances

In table 3 the correlation between coherence and cohesion ratings and cohesion ratio of the short performances is shown.

Table 3: Correlation between coherence and cohesion ratings and cohesion ratio of the short performances

Correlations

		COH_Fair. Avge_short	cohesion ratio short
COH_Fair.Avge_short	Pearson Correlation	1,000	,190
	Sig. (2-tailed)		,315
	N	30	30
cohesion ratio short	Pearson Correlation	,190	1,000
	Sig. (2-tailed)	,315	
	N	30	30

The hypothesis that the cohesion ratio functions as a valid predictor for the coherence and cohesion ratings is falsified since there is no significant correlation between the two variables.

Average sentence length in words was chosen as a predictor for the ratings of the grammatical competence. As can be seen from the output table below, there is no significant relationship between these two variables in the short performances. Therefore, the hypothesis that average sentence length predicts the ratings of grammar in the short performances is falsified.

Table 4: Correlation between grammar ratings and average sentence length in words of the short performances

Correlations

		GR_Fair. Avge_short	average sentence length in words
GR_Fair.Avge_short	Pearson Correlation	1,000	-,075
	Sig. (2-tailed)		,693
	N	30	30
average sentence length in words	Pearson Correlation	-,075	1,000
	Sig. (2-tailed)	,693	
	N	30	30

In the short performance there is only one predictor variable, namely *standardized type token ratio on the basis of 40* (S*TTR40*), which predicts the vocabulary ratings. The results (table 5) show that there is a moderate correlation between the *STTR40* and the vocabulary ratings from the short performances (0.373.), but the significance is only at the 0.05 level. All the other variables that were expected to act as predictors for the vocabulary rating, do not show significant correlation coefficients.

Table 5: Correlations between vocabulary rating and predictors for vocabulary of the short performances

Correlations

		VOC_Fair. Avge_short	standardised type token ratio basis 40	average word length in characters on the basis of types	average word length in characters on the basis of tokens	average word frequency in BNC based on tokens	average word frequency in BNC based on types
VOC_Fair.Avge_short	Pearson Correlation	1,000	,373*	,168	,003	-,259	-,199
	Sig. (2-tailed)		,042	,374	,986	,167	,291
	N	30	30	30	30	30	30
standardised type token ratio basis 40	Pearson Correlation	,373*	1,000	,068	,048	-,346	-,271
	Sig. (2-tailed)	,042		,719	,800	,061	,147
	N	30	30	30	30	30	30
average word length in characters on the basis of types	Pearson Correlation	,168	,068	1,000	,378*	-,344	,024
	Sig. (2-tailed)	,374	,719		,039	,063	,900
	N	30	30	30	30	30	30
average word length in characters on the basis of tokens	Pearson Correlation	,003	,048	,378*	1,000	-,063	,197
	Sig. (2-tailed)	,986	,800	,039		,742	,298
	N	30	30	30	30	30	30
average word frequency in BNC based on tokens	Pearson Correlation	-,259	-,346	-,344	-,063	1,000	,514**
	Sig. (2-tailed)	,167	,061	,063	,742		,004
	N	30	30	30	30	30	30
average word frequency in BNC based on types	Pearson Correlation	-,199	-,271	,024	,197	,514**	1,000
	Sig. (2-tailed)	,291	,147	,900	,298	,004	
	N	30	30	30	30	30	30

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

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

Apart from the correlation between the vocabulary rating and the *STTR40*, there are two other significant correlations illustrated in table 5. On the one hand there is a moderate correlation of 0.378 at a 0.05 significance level between the *average word length on the basis of types* and the *average word length on the basis of tokens*. Further, the table indicates a medium correlation of 0.514 that is highly significant at a 0.01 level between *average word frequency in the BNC on the basis of types* and *average word frequency in the BNC on the basis of tokens*.

5.1.2.2. Long performances

In contrast to the short performances, the predictor variables performed better in the long performances. A majority of the predictor variables show some relationship with the human ratings. One variable which does not predict the ratings is the cohesion ratio. As can be seen from table 6, there is no significant correlation (0.193) between the coherence and cohesion measure and the predictor variable *cohesion ratio* in the long performances. Therefore the hypothesis can be falsified.

Table 6: Correlation between coherence and cohesion ratings and cohesion ratio in the long performances

Correlations

		COH_Fair. Avge_long	cohesion ratio long
COH_Fair.Avge_long	Pearson Correlation	1,000	,193
	Sig. (2-tailed)		,306
	N	30	30
cohesion ratio long	Pearson Correlation	,193	1,000
	Sig. (2-tailed)	,306	
	N	30	30

For predicting the ratings of the test takers' grammatical competence the predictor variable *average sentence length in words* was chosen. The underlying hypothesis was that longer sentences are more complex in syntax and therefore elicit a higher rating. Correlation statistics though illustrate (Table 7) that this hypothesis is not true. There is no significant correlation (-0.190) between the human grammar ratings and *average sentence length in words*.

Table 7: Correlation between grammar ratings and average sentence length in words in the long performances

Correlations

		GR_Fair. Avge_long	average sentence length in words
GR_Fair.Avge_long	Pearson Correlation	1,000	-,190
	Sig. (2-tailed)		,314
	N	30	30
average sentence length in words	Pearson Correlation	-,190	1,000
	Sig. (2-tailed)	,314	
	N	30	30

The rating dimension which can be best predicted in the long performances is vocabulary. Two out of five predictor variables performed as expected. As illustrated in table 8, the highest significant correlation is found between the vocabulary rating and the *average word frequency in the BNC based on types* (-0.637), followed by *average word length in characters on the basis of tokens* (0.399). The latter correlation is of moderate strength and it is only significant at a 0.05 level whereas the first has a significance level of 0.01.

Table 8: Correlations between the vocabulary ratings and the predictor variables (STTR40, average word length in characters on the basis of types and on the basis of tokens, average word frequency in the BNC on the basis of types and on the basis of tokens)

Correlations

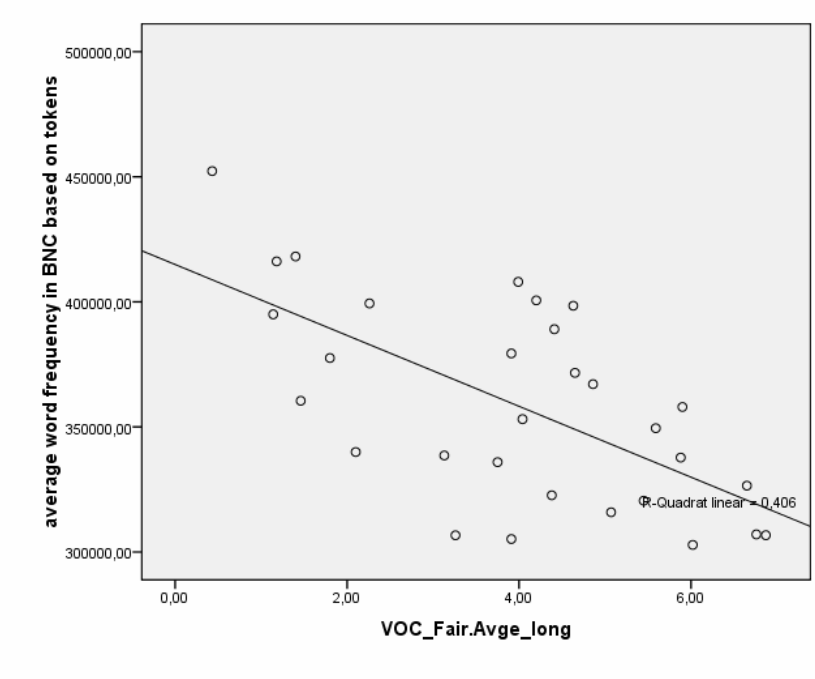
		VOC_Fair.Avge_long	standardised type token ratio on the basis of 40	average word length in characters on the basis of types	average word length in characters on the basis of tokens	average word frequency in BNC based on tokens	average word frequency in BNC based on types
VOC_Fair.Avge_long	Pearson Correlation	1,000	,338	,293	,399*	-,637**	-,268
	Sig. (2-tailed)		,068	,117	,029	,000	,153
	N	30	30	30	30	30	30
standardised type token ratio on the basis of 40	Pearson Correlation	,338	1,000	,230	,136	-,370*	-,330
	Sig. (2-tailed)	,068		,222	,475	,044	,074
	N	30	30	30	30	30	30
average word length in characters on the basis of types	Pearson Correlation	,293	,230	1,000	,611**	-,004	-,017
	Sig. (2-tailed)	,117	,222		,000	,981	,927
	N	30	30	30	30	30	30
average word length in characters on the basis of tokens	Pearson Correlation	,399*	,136	,611**	1,000	-,274	,059
	Sig. (2-tailed)	,029	,475	,000		,144	,756
	N	30	30	30	30	30	30
average word frequency in BNC based on tokens	Pearson Correlation	-,637**	-,370*	-,004	-,274	1,000	,452*
	Sig. (2-tailed)	,000	,044	,981	,144		,012
	N	30	30	30	30	30	30
average word frequency in BNC based on types	Pearson Correlation	-,268	-,330	-,017	,059	,452*	1,000
	Sig. (2-tailed)	,153	,074	,927	,756	,012	
	N	30	30	30	30	30	30

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

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

As can be seen from table 8 and the scatterplot (Figure 1) the correlation between the vocabulary ratings and the *average word frequency in the BNC based on types* is a negative one. This means that test takers who used many frequent words from the BNC, got a lower rating for vocabulary. The most frequent words in the BNC are *the, of, and, a*. These very frequent words are learned at an early stage in second language acquisition. From the correlation analysis we might conclude that the more infrequent words from the BNC are used in the E8 writing performances the higher is the vocabulary rating. Statistical evidence is illustrated in the scatterplot which shows a downward line from the left to the right suggesting negative correlation, and in the correlation coefficient (-0.637).

Figure 1: Scatterplot: Vocabulary fair measure and average word frequency in the BNC based on the tokens in the long performances



From the statistical analysis (Table 8) it can be seen that there is a weak correlation (0.338) between the vocabulary ratings and the *STTR40* in the long performances. This correlation may be only interpreted as a tendency towards a relationship between these two variables.

Both, *average word length on the basis of types* (0.293) and *average word frequency in the BNC on the basis of types* (-0.268) are not significantly related to the vocabulary ratings of the long performances.

In addition to the statistical analyses described above, all predictor variables were correlated with all rating dimensions, in order to investigate whether certain predictor variables predicted human ratings from other dimensions than expected.

Table 9: Correlations between all rating dimensions all predictor variables of the long performances

Correlations

		standardised type token ratio on the basis of 40	average word length in characters on the basis of types	average word length in characters on the basis of tokens	average word frequency in BNC based on tokens	average word frequency in BNC based on types	average sentence length in words	cohesion ratio long
TA_Fair.Avge_long	Pearson Correlation	,379*	,162	,170	-,462*	-,267	-,006	,022
	Sig. (2-tailed)	,039	,391	,369	,010	,153	,975	,910
	N	30	30	30	30	30	30	30
COH_Fair.Avge_long	Pearson Correlation	,210	,317	,336	-,632**	-,254	-,002	,193
	Sig. (2-tailed)	,265	,088	,069	,000	,176	,992	,306
	N	30	30	30	30	30	30	30
GR_Fair.Avge_long	Pearson Correlation	,291	,297	,523**	-,607**	-,156	-,190	,071
	Sig. (2-tailed)	,119	,111	,003	,000	,411	,314	,707
	N	30	30	30	30	30	30	30
VOC_Fair.Avge_long	Pearson Correlation	,338	,293	,399*	-,637**	-,268	-,070	,125
	Sig. (2-tailed)	,068	,117	,029	,000	,153	,712	,511
	N	30	30	30	30	30	30	30

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

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

As this table outlines, there are predictor variables that predict rating dimensions that were not primarily expected. From the output in table 9 it can be seen that *STTR40* shows a low correlation (0.379) at the 0.05 significance level with task achievement. Since this study did not focus on investigating any predictor variable for the task achievement ratings, this weak correlation should be focused on in further research.

Furthermore, the output of this correlation analysis demonstrates that *average word length in characters on the basis of tokens* does not only predict the vocabulary ratings, as it was hypothesised, but it also predicts the grammar ratings. There is a moderate and highly significant correlation (0.523).

Average word frequency in the BNC based on tokens is a predictor variable that predicts all four rating dimensions, showing moderate to high correlations throughout. The correlation coefficients range from -0.462 with the task achievement ratings to -0.637 with the vocabulary ratings.

5.2. Results of regression analysis

From the results above it can be seen that there are some relationships between the human ratings and the predictor variables in the long performances. In order to explore the correlations between the continuous dependent variables - in this case the human ratings - and a set of independent variables - in this case the predictor variables - in a more sophisticated way, a regression analysis is used.

From correlation analysis we already know that the ratings for task achievement correlate with the predictor variables of *STTR40* (0.379) and *average word frequency in the BNC based on tokens* (-0.462). These two predictor variables are entered into the multiple regression analysis.

Table 10: Model Summary task achievement ratings in the long performances

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,514[a]	,264	,209	2,03565

a. Predictors: (Constant), average word frequency in BNC based on tokens, standardised type token ratio on the basis of 40

b. Dependent Variable: TA_Fair.Avge_long

This table from regression analysis demonstrates that 26.4 per cent (the value *R Square* 0.264 multiplied by 100) of the variance in the dependent variable (TA_Fair_Avge_long – task achievement fair measure) is explained by the model which includes the variables *of average word frequency in the BNC based on tokens,* and *STTR40*.

Table 11: Coefficients: task achievement ratings, STTR40 and average word frequency in the BNC based on tokens

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1,987	9,296		,214	,832	-17,087	21,061					
	standardised type token ratio on the basis of 40	,129	,095	,241	1,355	,187	-,066	,324	,379	,252	,224	,863	1,159
	average word frequency in BNC based on tokens	-2,105E-5	,000	-,373	-2,099	,045	,000	,000	-,462	-,375	-,347	,863	1,159

a. Dependent Variable: TA_Fair.Avge_long

In Table 11 the **Beta** value (standardized coefficients) tells us that there is a positive relationship between the predictor variable *STTR40* and the task achievement ratings in the long performances (0.241). In other words, the variable *STTR40* contributed to the prediction of the dependent variable, the task achievement ratings. Further this means that as the *STTR40* increases the ratings for task achievement increase.

In the case of the predictor variable *average word frequency in the BNC based on tokens*, there is a very weak negative relationship with the task achievement ratings (-2.105E-5).

The correlation analysis in the long performances shows a significant correlation between the coherence and cohesion ratings and the *average word frequency in the BNC based on tokens* (-0.632). A regression analysis is carried out in order to investigate this relationship in a more sophisticated way.

Table 12: Model Summary coherence and cohesion ratings and average word frequency in the BNC based on tokens in the long performances

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,632[a]	,400	,378	1,33770

a. Predictors: (Constant), average word frequency in BNC based on tokens

b. Dependent Variable: COH_Fair.Avge_long

The model summary in table 12 demonstrates that the predictor *average word frequency in the BNC based on tokens* accounts for 40 per cent (R Square 0.400) of the variance in the coherence and cohesion ratings.

Table 13: Coefficients: coherence and cohesion ratings of the long performances and average word frequency in the BNC based on tokens

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	13,292	2,208		6,019	,000	8,768	17,815					
	average word frequency in BNC based on tokens	-2,641E-5	,000	-,632	-4,316	,000	,000	,000	-,632	-,632	-,632	1,000	1,000

a. Dependent Variable: COH_Fair.Avge_long

Table 13 clearly shows that the predictor variable has a negative **Beta** value. The beta value (-0.632) indicates that the predictor variable makes a strong, significant (Sig. 0.000) contribution to explaining the dependent variable.

As correlation analysis has shown highly significant correlations between the human grammar ratings, *average word frequency in the BNC based on tokens* (-0.607) and *average word length in characters on the basis of t*okens (0.523), a regression analyses is made to investigate the relationship between these variables in a more sophisticated way.

Table 14: Model Summary grammar ratings and average word frequency in the BNC based on tokens and average word length in characters on the basis of tokens in the long performances

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,711[a]	,506	,469	1,31076

a. Predictors: (Constant), average word frequency in BNC based on tokens, average word length in characters on the basis of tokens

b. Dependent Variable: GR_Fair.Avge_long

As demonstrated in Table 14, the variance in the dependent variable is explained by the model by 50.6 per cent (R Square 0.506), including the predictor variables of *average word length in characters based on tokens* and *average word frequency in the BNC based on tokens* at significance levels of 0.011 and 0.001.

Table 15: Coefficients: grammar ratings and average word frequency in the BNC based on tokens and average word length in characters on the basis of tokens.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-3,557	6,507		-,547	,589	-16,908	9,795					
	average word length in characters on the basis of tokens	3,578	1,303	,386	2,746	,011	,905	6,252	,523	,467	,371	,925	1,081
	average word frequency in BNC based on tokens	-2,221E-5	,000	-,501	-3,564	,001	,000	,000	-,607	-,566	-,482	,925	1,081

a. Dependent Variable: GR_Fair.Avge_long

The **b** value in Table 15 shows that there is a positive relationship between the constant and the *average word length in characters on the basis of tokens*, whereas there is a negative relationship between the constant and the second predictor variable *average word frequency in BNC based on tokens*. Both predictor variables are significant. The largest **Beta** coefficient in Table 15 is -0.501 (ignoring the negative sign out the front), which is for *average word frequency in BNC based on tokens.* This means that this variable makes the strongest unique contribution to explaining the dependent variable, grammar ratings.

Table 16: Model Summary vocabulary ratings and average word frequency in the BNC based on tokens and average word length in characters on the basis of tokens in the long performances

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,679[a]	,461	,421	1,38985

a. Predictors: (Constant), average word frequency in BNC based on tokens, average word length in characters on the basis of tokens

b. Dependent Variable: VOC_Fair.Avge_long

In the model summary in Table 16 we can see that 46.1 per cent (R Square 0.461) of the variance is explained by the model including two predictor variables (*average word frequency in the BNC based on tokens* and *average word length in characters on the basis of tokens*).

Table 17: Coefficients: vocabulary ratings and average word frequency in the BNC based on tokens and average word length in characters on the basis of tokens.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3,510	6,900		,509	,615	-10,647	17,667					
	average word length in characters on the basis of tokens	2,281	1,382	,243	1,651	,110	-,554	5,116	,399	,303	,233	,925	1,081
	average word frequency in BNC based on tokens	-2,569E-5	,000	-,571	-3,886	,001	,000	,000	-,637	-,599	-,549	,925	1,081

a. Dependent Variable: VOC_Fair.Avge_long

The **b** value illustrates a positive relationship between the dependent variable (vocabulary ratings of the long performances) and one of the predictor variables (*average word length in characters on the basis of tokens*). There is a negative relationship between the dependent variable and the second predictor variable (*average word frequency in the BNC based on tokens*). The latter predictor variable is significant at a 0.001 level whereas the first one shows no significance. The **Beta** coefficient for *average word frequency in the BNC based on t*okens is -0.571. This means that this variable makes the strongest unique contribution to explaining the dependent variable.

6. Discussion

The results of previous research differ in the short and long performances. On the one hand it was found that the correlations among the rating dimensions in the short performances were slightly lower than in the long performances. The vocabulary dimension could be seen as the predicting rating dimension for the other three dimensions task achievement, coherence and cohesion and grammar as the vocabulary ratings correlated significantly with all dimensions.

Furthermore, it was found out that *STTR* is the only statistically significant predictor variable for the short performances. The other variables, which were expected to predict the ratings in the short performances, did not show any relationship to the ratings. We may conclude that raters assess different things when rating the short performances as the rating dimensions did not show significant correlations among the rating dimensions. Regarding the predictor variables, the analyses illustrate that only the *STTR* has a significant relationship to the vocabulary ratings. The other predictor variables were not significantly related to the ratings and therefore further research is needed to find out what raters focus on when rating the short performances.

In the long performances the analyses demonstrate that all rating dimensions correlate among each other. From this result, it can be concluded that the E8 Writing Test measures one construct, namely the competence of writing in terms of task achievement, coherence and cohesion, grammar and vocabulary. Some of the predictor variables (*STTR, average word length in characters on the basis of tokens, average word frequency in the BNC based on tokens*) performed as it was hypothesised.

Further, it was found out that some predictor variables were good indicators for rating dimensions that they were not expected to predict. The most outstanding predictor variable is the *average word frequency in the BNC based on the tokens*, as it statistically pretends to be related to all four rating dimensions in the long performances. Overall it can be said, that there were more textual features found which tend to predict the ratings in the long performances.

When it comes to predicting ratings in writing performances of different length, there is plenty of room for further research. One could concentrate on the differences

between short and long performances in connection with the predictor variables. Moreover, further research could investigate on more features and variables which might be able to predict the four rating categories.

References

Alderson, J.C. and C. Clapham, D. Wall. 1996. *Language Test Construction and Evaluation*. Oxford: Oxford University Press.

Bachman, L.F. 1991. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F. and A.S. Palmer. 1996. *Language Testing in Practice.* Oxford: Oxford University Press.

Davies, A. et al. 1996. *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Frey, E. and H.J. Heringer. 2007. "Automatische Bewertung schriftlicher Lernerproduktion". Linguistische Berichte 11, 331-345. Hamburg: Helmut Buske Verlag.

Gassner, O., Mewald, C. and G. Sigott. 2008. Testing Writing. Specifications for the E8-Standards Writing Tests. LTC Technical Report 4. Wien: bm:ukk. Available as download from: http://www.uni-klu.ac.at/ltc/downloads/LTC_Technical_Report_4.pdf

Hughes, A. 2007. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Scott, M. 2004. *WordSmith Tools* version 4. Oxford: Oxford University Press.

Shaw, S.D. and Weir C.J. 2007. *Examining Writing. Research and practice in second language writing.* Cambridge: Cambridge University Press.

Sigott, G. 2004. *Towards Identifying the C-test Construct*. In: R. Grotjahn and G.Sigott (Hrsg.): Language Testing and Evaluation. Frankfurt am Main; Wien: Peter Lang.

Weigle, S.C. 1998. "Using FACETS to model rater training effects." *Language Testing 15/2,* 263-287.

Weigle, S.C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J.: *Language Testing and Validation*. 2005. An Evidence-Based Approach. Basingstoke: Palgrave Mcmillan.

Internet sources

Davies, Mark. (2004) BYU-BNC: The British National Corpus. Available online at http://corpus.byu.edu/bnc (accessed 10 February 2009).