Variation in Assessment

A Coh-Metrix Analysis of Evaluation of Written English in the Swedish Upper Secondary School

Hugo Simion Malmcrona
Variation in Assessment

A Coh-Metrix Analysis of Evaluation of Written English in the Swedish Upper Secondary School

Hugo Simion Malmcrona

Abstract

Reliable evaluation is an important part of language education. However, reliable evaluation of student writing is notoriously difficult to achieve nationally. To better understand what evaluators value in upper secondary English education in Sweden this study has examined correlations between grading and linguistic variables in student writing using Coh-Metrix, a natural language processor. Previous studies in Hong Kong and the U.S. have shown that linguistic and lexical sophistication have a high correlation with good grades, while explicit cohesive devices have a negative correlation with grades. Results have varied depending on level of proficiency and there are indications that evaluators have cultural differences. Seventy-nine essays from the national test for the course English 6, from two different schools and teachers were analysed in this study. The study asked: What language variables correlate with grades, and how do two teachers differ in their value of quality? In order to answer this, correlations between grades and all 106 Coh-Metrix indices were calculated and t-tested for significance. The ten most highly correlating variables were then compiled, with very similar indices excluded. Correlation scores of the same variables but for both teachers separately were then compiled for comparison. The ten highest correlations for both teachers separately were also compiled along with scores for the same indices for the other teacher. The results showed similarities with previous studies. Lexical sophistication correlated significantly with grades and indices measuring explicit cohesive devices showed significant negative correlations with grades. Surprisingly however, syntactic complexity did not correlate clearly with grades. The highest correlating indices differed between teachers but were mostly connected to cohesion or word choice. In all, the results showed that while teachers seem to differ in their evaluative focus, they do in fact value similar aspects of student writing.

Keywords

Writing assessment, lexical sophistication, cohesion, Coh-Metrix, L2 writing.
Contents

1. Introduction .................................................................................................................. 1
   1.1 Literature review ..................................................................................................... 1
       1.1.1 NLP studies of student writing ........................................................................ 2
   1.2 Aim and research questions .................................................................................... 4
2. Methodology .................................................................................................................. 5
   2.1 The Swedish national test ....................................................................................... 5
   2.2 Coh-Metrix .............................................................................................................. 5
   2.3 Procedure ............................................................................................................... 6
3. Results .......................................................................................................................... 7
   3.1 All essays ................................................................................................................ 7
       3.1.1 Lexical sophistication ..................................................................................... 8
       3.1.2 Explicit cohesive devices ............................................................................... 9
   3.2 Group A .................................................................................................................. 11
       3.2.1 Explicit cohesive devices .............................................................................. 11
       3.2.2 Lexical sophistication .................................................................................... 12
       3.2.3 Syntactic complexity ..................................................................................... 13
   3.3 Group B .................................................................................................................. 14
   3.4 Concluding comparison of group A and B ............................................................. 15
4. Discussion and conclusion ............................................................................................ 15
   4.1 The implications of this study for the Swedish school ............................................ 15
   4.2 Implications for field of research .......................................................................... 17
   4.3 Conclusion .............................................................................................................. 18
References ......................................................................................................................... 19
1. Introduction

Reliable evaluation is a vital part of education. It is necessary for diagnostic purposes to indicate what teachers should focus on, and it is important for creating meaningful formative assessment. Since final grades are the most important determining factors for students’ university applications in Sweden, reliable grades are also essential for a fair application process. Yet, Swedish teachers’ approach to grading is varied (Vallberg Roth, Gunnemyr, Londos, & Lundahl, 2016) and the consistency of the grading of the national test has been shown to be lacking (Gustafsson, Cliffordson, & Erickson, 2014). The grading of the national test in Sweden is therefore an important area of research. In order to understand the range of variation between different graders, this study will examine differences in grading between two teachers. The analysis will examine what language features correlate with high grades, or in other words what they actually value when assessing texts.

The connection between linguistic features and human evaluation of writing has been studied by Crossley and McNamara using the natural language processor (NLP) Coh-Metrix (2011; 2012). Their results showed that lexical diversity, usage of low frequency words and complex syntax were connected with high quality writing, while frequent cohesive devices were more common in lower rated essays. In other words, linguistic sophistication rather than explicit cohesive devices is premiered by evaluators (Crossley & McNamara, 2011; 2012). Further, writers may use different techniques to show their writing strengths and hide their weaker areas. This means that good writing can be rated highly due to different types of variables. Text length, however, consistently shows a clear correlation with high grades (Jarvis, Grant, Bikowski & Ferris, 2003).

Gustafsson et al. (2014) based their discussion of the national test on a study that compared teachers’ grading of their own students with external graders. This showed a lack of grader reliability, which was especially poor for essays. However, they could not further explain what teachers value when they assign grades. The approach of using NLP to study the national test thus can offer interesting new insights into what evaluators regard as high quality writing in English education in Sweden. By understanding what assessors focus on and how they differ, a clearer picture of grading inconsistencies may emerge. The present study will therefore analyse graded student writing from the national test for the course English 6 in upper secondary school. Language variables that correlate with grades will be examined and the results from two different groups of students will be compared.

1.1 Literature review

Validity and reliability are cornerstones of assessment. They are often dichotomous; when reliability is high validity is low (Lundahl, 2012). For example, a fill in the gap test is highly reliable, because many different graders will award the same score, but if the test is supposed to measure writing proficiency it has low validity. An essay on the other hand would have high validity, but different graders would award a varied score, meaning poor reliability. This dilemma has been identified in the grading of the national test, where the
short answer test was graded more reliably than the essay (Gustafsson et al., 2014). This has led to a discussion on how to merge high reliability and validity. Proponents argue that the only fair way to assess writing is to follow the same criteria and train assessors to grade equally, while others argue that it is impossible and assessment should instead focus on pedagogy and how students can improve (Casanave, 2004). Elbow (1993) for example, argues that reliability is so poor that grading is never fair, and should therefore be avoided as far as possible. Focus should instead be on what would today probably be called formative assessment. That being said, grades are important in the current school system in Sweden and should therefore be as fair as possible. Because grades are used for university and upper secondary school applications they cannot only be used for pedagogical purposes.

In a classic study from 1912, 142 high school teachers graded two different texts (Starch & Elliot, 1912). The researchers were surprised by the wide range of marks and noted that 22 of the 142 teachers did not give a passing grade to one essay. The study also showed that teachers were not consistent in their grading and that the poorer essay got much less consistent points. This study was replicated more recently. Seventy-three teachers from one U.S. school district trained in the same assessment method graded a single paper. They awarded the student essay scores between 50 to 96 on a 100-point scale, with a standard deviation of 9.6. In letter grades, that meant the same essay received an A by ten graders, a B by 18, a C by 30, a D by nine and finally an F by six graders (Brimi, 2011). This study clearly shows the problems of implementing reliable grading in real schools. Similarly, the national test in Sweden, which is supposed to ensure equal grading across the country is not graded sufficiently reliably (Vallberg Roth, et al., 2016). Student compositions in the national test are graded more generously by their teachers than by external evaluators, and course grades also see inflation (Gustafsson et al., 2014).

An approach to the lack of reliability in grading student writing is automated essay scoring (AES). Already in 2003 one such system was shown to correlate as highly with human graders as human graders do (Landauer, Laham & Foltz, 2003). However, AES has been shown to be less accurate when grading poor and very good writing (Tsai, 2012). Further, while AES correlates with human graders and is naturally highly reliable, it is less valid (Cohen, Levi & Ben-Simon, 2018). This means AES should not be used on its own for grading but can serve as a good tool for graders. The same technology also enables researchers to study texts in new ways. One of these natural language processors (NLP), Coh-Metrix, has been used in several studies to analyse student writing.

1.1.1 NLP studies of student writing

Crossley and McNamara (2012) analysed essays by Hong Kong high school students using Coh-Metrix. The variables that correlated with grades were used in a regression analysis to see if they predicted human scoring. The results showed that five variables significantly predicted high grades: lexical diversity, word frequency, word meaningfulness, aspect repetition and word familiarity. This meant that highly rated essays were not more cohesive, but rather more linguistically sophisticated. However, low cohesion does not mean that coherence was low. Cohesion in this context refers to explicit cohesive devices that can be measured in a text, while coherence is the mental
process occurring when reading. In another article, Crossley and McNamara (2011) reiterated that similar results were true for L1 writing, but they also showed that essays with low cohesion and high linguistic complexity were judged to be the most coherent. They hypothesised that low cohesion texts make readers interact with texts more, thus increasing coherence.

Based on Crossley and McNamara’s (2012) study, a Swedish teaching degree project examined grading in adult English education at an upper secondary school level in Sweden (Westerlund, 2019). Linguistic sophistication played a similar role, but there was no correlation between cohesion and grades in the student essays analysed. However, the study also analysed example texts provided for evaluators as a grading guide, and found a negative correlation between explicit cohesive devices and grades in these texts (Westerlund, 2019). The importance of lexical and syntactical sophistication was further shown in a study of essays written by undergraduate college students in the U.S. All essays were evaluated by five different English tutors with at least one year of experience. Three Coh-Metrix indices were identified to predict grades most: syntactic complexity (number of words before main verb), word frequency (Celex, logarithm for all words), and lexical diversity (Measure of Textual Lexical Diversity). Like the student essays in Westerlund’s (2019) study, no measures of cohesion correlated significantly with grades. (McNamara, Crossley & McCarthy, 2010a).

The suggestion of a negative relationship between perceived quality and cohesion was further nuanced by McNamara, Crossley and Roscoe (2013) who concluded that while quality essays had low content word overlap and did not have higher overlap between sentences, “they did have some sense of global, semantic cohesion”, as manifested by a positive relationship between LSA given new and quality (2013, p. 511). This idea has consequently been researched, and findings have indicated that, for expert raters, global cohesion, not local cohesion, creates coherence and is indicative of quality (Crossley, Kyle & McNamara, 2016).

Even more differing results were presented in a recent study. Essays by U.S. college students were examined, but from basic writing classes graded by graduate students. As in the previously mentioned studies lexical complexity correlated with quality but syntactic complexity correlated negatively with quality, as opposed to the results discussed above. Another difference from earlier research was a significant correlation between quality and referential cohesion (MacArthur, Jennings & Philippakos, 2019). The lower proficiency level of these students might in part explain the difference. Another study of essays by low-skilled college students in the U.S. tested whether the results from McNamara et al. (2010a) applied to that group. Lexical diversity could to a limited degree predict quality, and word frequency and syntactic complexity did not differentiate between high and low graded essays in this lower level group. The researchers proposed that the low proficiency writers were simply more varied in their low quality. Further, the results also differed based on genre. The study was made on persuasive essays and written summaries and quality was predicted by different Coh-Metrix indices depending on genre. The persuasive essays were predicted by the referential cohesion index argument overlap in all sentences, binary, mean and the lexical diversity index type-token ratio, all words. The strongest predictors of quality of the written summary were the referential cohesion measure content word overlap, adjacent sentences, proportional, standard
Coh-Metrix has also been used to predict performance of U.S. middle schoolers. Nine indices were used to measure three levels of language: discourse, sentence and word level. The results of two different state tests were examined and none were predicted by the word level. One was only predicted by discourse and the other by discourse and sentence level (Wilson, Roscoe & Ahmed, 2017). Writing quality thus seems to be predicted by different language variables in different levels of proficiency. The word level in Wilson et al. (2017) did not predict text quality but word frequency has, as discussed above, predicted quality in college writing. A study comparing essays by U.S. students in grade 9, 11 and college freshmen showed that complex syntax and vocabulary increased with grade level. Simultaneously, explicit cohesive devices decreased (Crossley, Weston, McLain Sullivan & McNamara 2011b). This clearly means different language features will predict essay quality depending on grade level.

Most of the studies discussed above were made on L1 speakers of English. It is therefore worthwhile to mention two studies of EFL students that studied cohesion, although not using Coh-Metrix. A study of Filipino college students showed no difference in cohesion between grades (Castro, 2004). A Spanish study on the other hand, counted conjunctions in texts by secondary education students and found a significant relationship between conjunction frequency and essay score. Use of conjunctions also increased from grade three to four (Martinez, 2015). This indicates that the connection between language variables and perceived quality by graders is differs over the world, making it fruitful to study this connection among Swedish students of English.

**1.2 Aim and research questions**

This study aims to further our understanding of how grading of the national exam in Sweden functions by examining linguistic features of student essays. This will hopefully give new insights into what teachers perceive as high- or low-quality writing, and thus further our understanding of grading in Sweden. This project studies what teachers value in essays by their own students and is therefore not simply interested in reliability, but rather how two teachers approach and value student writing. It will, furthermore, broaden the field of research concerned with text evaluation and linguistic features. The study asks the following research questions:

- Which Coh-Metrix language variables correlate with grades?
- How do two teachers differ in their value of quality, as expressed by a Coh-Metrix analysis?
2. Methodology

2.1 The Swedish national test

In Sweden, upper secondary school teachers give course grades based on the entire body of work a student has produced during a course. The national test is supposed to ensure that teachers make equal and fair evaluations of students’ abilities and to help them give fair final grades. The writing section is carried out simultaneously in all upper secondary schools across the country, digitally and with no access to spelling or grammar checking software, or to a dictionary. For students with special needs individual adjustments to the test are allowed, such as extra writing time, or use of voice synthesis (Skolverket, nationella prov). In this study essays from the writing part of the test for the course English 6 were analysed. All students were in year two of the three-year upper secondary school. Students were asked to write an argumentative or discussion essay between 300 to 600 words long in 100 minutes. Prior to the test they were given material on the topic of the test for preparation and inspiration. They were not allowed to bring this to the test. Every teacher is responsible for grading their class, but they are provided with extensive assessment guidelines such as graded example essays with explanatory comments and reasoning, as well as comments on how to interpret the knowledge requirements. All essays are anonymised prior to evaluation.

Even though the test is perceived to ensure reliable grading nationally, it is not effective enough to enable equal grading (Vallberg-Roth et al., 2016). Especially in the written part, where a student’s ability to produce a long text is assessed, there is often a discrepancy between teacher evaluation and external evaluation (Gustafsson et al., 2014). The grading of the test is therefore important and needs to be examined further.

2.2 Coh-Metrix

Coh-Metrix is an NLP that was originally developed to measure cohesion, hence the name, but has developed into a broader linguistic tool. Development was originally started in 2002 and Coh-Metrix 3.0 is now openly available at cohmetrix.com. An important definition for the tool is the difference between cohesion and coherence. This study follows the creators’ view that cohesion refers to observable, and therefore measurable, aspects of a text, while coherence occurs in the mind of the reader, albeit of course influenced by cohesion (McNamara, Graesser, McCarthy & Cai, 2014).

Language difficulty can be analysed on six levels: words, syntax, textbase, situation model, genre and rhetorical structure and finally, pragmatic communication. The first two are quite straightforward. Vocabulary affects ease of understanding, and complex and long sentences are more difficult to digest. The textbase level concerns ideas in the text without the surface code of vocabulary and syntax. The situation model refers to the subject matter, referential content or the microworld of events, people, objects etc., in a text. Text genre is self-explanatory and rhetorical structures concerns text organisation and discourse function, such as discerning between questions and beliefs. Finally, the pragmatic level is about the aims of the author, humour, irony and attitude. Coh-Metrix measures all of these levels in different ways, except for the pragmatic level. It is
unfortunately beyond the programme’s scope and is recognised as a limitation (Graesser & Forsyth, 2013; McNamara et al., 2014)

As the literature review in the previous section shows, Coh-Metrix has been used to great effect in many studies of essay quality, and was validated in a study where Coh-Metrix indices of cohesion better distinguished between high and low cohesion texts than previously common readability measures (McNamara, Louwerse, McCarthy & Graesser, 2010b). The current study’s use of Coh-Metrix is warranted by Crossley and McNamara’s (2011) conclusion that Coh Metrix “can help us better understand the interplay between linguistic features and human judgements of essay quality” (p. 187).

Values for 106 indices divided into 11 banks, or groups of indices, are given in Coh-Metrix 3.0. The indices used in this study will be explained as they appear, but are from these banks: descriptive, text easability principal component scores, referential cohesion, latent semantic analysis, lexical diversity, syntactic complexity, situation model, word information and readability (McNamara et al., 2014). Only indices among the ten highest correlations to grades for all essays, and group A and B separately were used, so indices from the remaining two banks were not included.

2.3 Procedure

Seventy-nine essays were collected for this study, all written in the spring of 2019. Forty essays were graded by one teacher in a school in the vicinity of Gothenburg (group A), and 39 essays graded by another teacher in the Stockholm region (group B). Both teachers are experienced English teachers from Sweden and the essays were written by their students. The essays ranged from around 270 to 600 words in length. There are some clear differences between the two groups. The average grade in Group A was 3.425 (if F=0 and A=5) but only 2.513 in B. The standard deviation shows that group B (std 1.745) had much more varied results than A (0.931). Table 1 shows all grades. Like Crossley and McNamara (2012), the number of essays per grade were not uniform. All essays were not graded by the same assessors, so this study does not explore how assessors differ in grading the same texts, but rather what two teachers value in their own students’ writings.

The Coh-Metrix data was compiled in Excel along with grades, which were converted to numericals where F=0 and A=5. Pearson correlations were then calculated for all essays as well as for groups A and B separately. The significance of correlations was then calculated using the two tailed Student’s t-distribution function. These p values were deemed significant if below 0.05, as is common (McNamara et al., 2014). The ten strongest correlations between variables and grades, for all essays, only group A and then only B were compiled. However, very similar indices were excluded in favour of the one
with the highest correlation. Thus, in Table 2 essay length is only represented by word count, not sentence count as well. Similarly, only one measure of lexical diversity is included per table. For uniformity z-scores, not percentiles were used, and standard deviation indices were excluded. Finally, the effect size of $r$ values should be noted. $r=0.1$ is considered small, 0.3 medium and 0.5 large (Larson-Hall, 2010).

3. Results

The results will be discussed in three sections. First, correlations between all essays and their grades will be presented. Second, correlations between grades and the essays in group A will be presented and compared to group B, before finally, the reverse is done. Unless self-explanatory, new Coh-Metrix indices will be explained as they appear.

3.1 All essays

Table 2 presents correlations ($r$ value) between Coh-Metrix variables and grades in all essays analysed. The corresponding correlation score of group A and B independently are

<table>
<thead>
<tr>
<th>Variable</th>
<th>All essays</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word count, number of words</td>
<td>$r$ value</td>
<td>$p$ value</td>
<td>$r$ value</td>
</tr>
<tr>
<td>0.494</td>
<td>&lt; 0.001</td>
<td>0.201</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>Hyponymy for nouns, mean</td>
<td>0.458</td>
<td>&lt; 0.001</td>
<td>0.278</td>
</tr>
<tr>
<td>Text Easability PC</td>
<td>-0.444</td>
<td>&lt; 0.001</td>
<td>-0.328</td>
</tr>
<tr>
<td>Referential cohesion, z score</td>
<td>-0.434</td>
<td>&lt; 0.001</td>
<td>-0.275</td>
</tr>
<tr>
<td>Content word overlap, all sentences, proportional, mean</td>
<td>0.428</td>
<td>&lt; 0.001</td>
<td>0.426</td>
</tr>
<tr>
<td>Lexical diversity, VOCD, all words</td>
<td>-0.395</td>
<td>&lt; 0.001</td>
<td>-0.170</td>
</tr>
<tr>
<td>Text Easability PC</td>
<td>-0.380</td>
<td>&lt; 0.001</td>
<td>-0.333</td>
</tr>
<tr>
<td>Narrativity, z score</td>
<td>-0.373</td>
<td>&lt; 0.001</td>
<td>-0.407</td>
</tr>
<tr>
<td>CELEX Log frequency for all words, mean</td>
<td>-0.365</td>
<td>&lt; 0.001</td>
<td>-0.486</td>
</tr>
<tr>
<td>LSA overlap, all sentences in paragraph, mean</td>
<td>0.374</td>
<td>&lt; 0.001</td>
<td>0.300</td>
</tr>
</tbody>
</table>
also shown for comparison. A $p$ value below 0.05 indicates that the result is significant. However, similar variables have been excluded, regardless of positive or negative correlation. Z-scores have been chosen instead of percentiles and standard deviation indices are not included. Further, because of inconsistent and ambiguous paragraphing and indentation in many essays, paragraph length is not included as a variable.

The highest correlation is word count, or the length of essays, which is in line with previous research (MacArthur et al., 2019). When correlations were calculated for group A only however, no significant results were found, while group B had a clear connection between essay length and grades. This can probably be explained by the differences between the two groups. The mean grade for group A was 3.4 and 2.5 for B, so that group B was on average about one grade level below A. The standard deviation also differed. It was 0.93 in group A, or about one grade level, but 1.75 in group B. Grades varied more in group B in other words. Further, in group B six essays got an F, while none in group A were failed. One of the F essays was even below the minimum word limit of 300. With that said, not all poor essays were very short; one failed essay was just two words shy of the 600-word maximum. The average length of essays in group B was 507 words and 542 for A, but more importantly for the results in the table, the standard deviation of word length in group A was 66 words and 101 for group B. This indicates that the greater the diversity of length and grades in a group, the greater the impact essay length has on grades.

### 3.1.1 Lexical sophistication

*Noun hypernymy* concerns how specific words are. A low value comes from use of general words such as *furniture*, rather than more specific words such as *stool* or *rocking chair* that give a higher value (McNamara et al., 2014). It is thus a measure of lexical sophistication. There is a high and significant correlation between noun hypernymy and grades. This is not surprising because, as has been discussed in the previous research section, measures of word choice have been shown to be highly predictive of perceived quality. With that said, the specific index result varies in the current study’s two groups as well as in the results of a previous study of 30 student essays from the national test for English 5. In that study noun hypernymy had a very high $r$ value of 0.844 (Westerlund, 2019). Group A in the current study had a substantially lower, but still significant, correlation score of 0.278. Group B on the other hand was in between the other scores with the strong correlation 0.523. Noun hypernymy is consistently linked with quality but to very different degrees in different groups. One possible explanations could be that noun hypernymy might have a bigger impact on grades if the essays are overall of a poorer quality. Group A received on average one grade higher than B, and the essays in Westerlund’s (2019) study had the same grade average as group B, but in the more basic course English 5 instead of 6 in the current study.

Consistent with previous studies *lexical diversity* showed a clear correlation with grades. Coh-Metrix has three different indices of lexical diversity, and another will be discussed in section 3.2. The *vocd*-index is calculated by samples of the ratio between unique words and number of words (so called type-token ratio, TTR) and ideal TTR curves. As it
measures the usage of unique words it indicates students’ lexical proficiency. However, it is also connected to cohesion because low lexical diversity means more word repetition and consequently, higher cohesion (McNamara et al., 2014). In Table 2 both groups have significant r values for this index. The same is true for the essays in Westerlund’s (2019) study, but not in the analysed example essays used as a grading guideline for teachers. Crossley and McNamara (2012) found a correlation of 0.426 which is identical to group A in table 2. Once again group B had a stronger correlation score of 0.622.

**CELEX Log frequency for all words, mean** reports how frequent the used words are in the English language. A logarithm is used because it better shows how word frequency is related to reading time (McNamara et al., 2014). In other words, just like lexical diversity this index both measures cohesion, or more specifically readability, and lexical sophistication, through the usage of rare words. It is therefore unsurprising that both groups showed significant negative correlations with grades, just like lexical diversity. Similar to that index, log frequency too had a lower r value in group A (-0.333) than B (-0.547). The student essays in Westerlund (2019) showed a correlation of -0.481, while the example essays had an r value of -0.714. That log frequency plays an important role in perceived essay quality seems clear, but what causes these diverging results is unclear. Once again, the lower r value in group A is plausibly explained by the less diverse grading. Nonetheless, this index reinforces the view that texts with higher readability are deemed of lower quality, while a more advanced lexical choice is connected to perceived high quality. The results are in line with McNamara et al. (2010a) where CELEX log frequency best discriminated between low- and high-quality essays.

The importance of word choice is demonstrated further by the self-explanatory index **word length**. In Table 2, word length has the lowest correlation (0.374) but it is still significant. Having said that, the result was not significant for group A in isolation (p-value 0.060). Again, because the general quality level was higher, the use of more advanced words was probably not as noteworthy as in group B where six students failed the assignment. There, word length had a significant r value of 0.465. The use of longer words thus seems to be more important in a group with lower grades. It is also possible that teacher B simply values sophisticated word choice more than teacher A.

### 3.1.2 Explicit cohesive devices

The text easability index **referential cohesion** measures how words and ideas overlap in a text, connecting ideas in the text for the reader (McNamara et al., 2014). For all essays a significant negative correlation of -0.444 was found, and the result is also significant for the two groups in isolation. Again, group B had a stronger r value than A. These results show clearly that repetition of words and ideas are more frequent in essays that teachers deem of lower quality. While the same index has not been shown to correlate with quality in the research discussed above, other cohesion variables have. However, MacArthur et al. (2019) found that referential cohesion (not the same Coh-Metrix variable) was positively correlated with quality. Another text easability index in the ten highest correlations was **narrativity**. Narrative text is similar to oral conversation and everyday language (McNamara et al., 2014). Group A showed no significant result for this index, but B showed a strong significant negative correlation of -0.512. High
narrativity is not suited for formal language and a likely reason for the difference between the two groups is the higher number of poorly rated essays in group B. To get a good grade on the national test “adaptation to purpose, recipient and situation” is needed (Skolverket, English). If poorer essays failed in this regard it is natural that the group with more low grades showed a distinct correlation between grades and narrativity. A negative relationship between essay score and narrativity has been identified previously. McNamara et al. (2013) found a more moderate, but still significant negative correlation of -0.222. They concluded that higher quality essays were more informational than narrative.

Content word overlap, all sentences, mean measures the proportional overlap of words between sentences. There was a clear difference between the two groups. A showed no significant results, while B had a strong significant r value of -0.581. The student essays in Westerlund’s study did not show a significant correlation to content word overlap but the example texts did, with an r value of -0.777 (2019). This index thus seems to vary drastically from different groups. However, seen as a measure of explicit cohesive devices, the significant result of a negative correlation with grades for all essays supports the notion that cohesion is connected to poorer essays.

Finally, two indices of Latent Semantic Analysis (LSA) must be discussed. LSA measures cohesion and “considers semantic overlap between explicit words and words that are implicitly similar or related in meaning” (McNamara et al., 2014, p. 66). LSA overlap, all sentences in paragraph measures semantic overlap in a full paragraph. The correlation score for all essays was -0.373, in line with the idea that cohesion is negatively correlated with grades. Interestingly however, this was the only non-significant (p-value 0.051) correlation score for group B in Table 2. Group A, on the other hand, had a significant correlation value of -0.407. This is not only different from group B, but also Westerlund (2019) who did not find such a strong r value, nor Crossley and McNamara (2012). LSA given/new was more consistent with significant scores for both groups: A (r value -0.486) and B (r value -0.370). Again, A had the higher correlation score indicating that semantic cohesion was more important for teacher A when grading. LSA given/new measures how much new information exists in a sentence. When the score is lower, more new information is introduced in sentences and cohesion is lower (McNamara et al., 2014). This means that less repetition and more new ideas lower the score. It is therefore understandable that it is negatively correlated to grades. In Wilson et al. (2017) the given/new score increased slightly from grade 6 to 8, indicating that the result would carry more weight, and increase correlation to grades, in higher proficiency groups. However, the increase was small, and many factors could potentially have affected the differences in group A and B. In Crossley and McNamara (2012) given/new had a grade to variable correlation of -0.265. No significant correlation was found in Westerlund’s (2019) student set, but the example essays showed a strong correlation of -0.739. Conversely, McNamara et al. (2013) found a small but significant positive correlation with LSA given/new and essay score. As with many indices results vary from different groups but LSA given/new seems to be an important part, among others, of mapping the importance of explicit cohesive devices in student writing proficiency.
3.2 Group A

Table 3 below lists the ten highest correlations between essay grade and Coh-Metrix indices for group A in isolation. The results of those indices for group B are presented in comparison. All results were significant for group A, but half were not in group B, indicating quite clear differences.

3.2.1 Explicit cohesive devices

Following Larson-Hall’s (2010) effect sizes, no \( r \) values reached the 0.5 threshold for a strong correlation, in group A, but \( LSA \) given/new was quite close at -0.486. That index has already been discussed above, but it is worth pointing out that semantic overlap showed the strongest correlation in group A but was not in the top ten for group B, as can be seen in table 4 in the next section. \( LSA \) overlap all sentences in paragraph has also been discussed in the previous section and it should just be noted that it was among the highest correlations for group A as well. One more LSA index was however among the top correlations only for group A: \( LSA \) overlap, adjacent sentences. This index only measures semantic overlap between adjacent sentences, instead of overlap between all sentences in a paragraph. As is clear in Table 3, they show similar values, although all sentences had a higher \( r \) value of -0.407 compared to -0.387 for adjacent sentences. No significant results were found in group B for these two LSA indices. As has been mentioned in the previous section, \( LSA \) given/new did correlate with grades for group B. So, less given information in a text correlated with higher grades in both groups, while the repetition of semantically similar words was only negatively correlated with grades in group A.

### Table 3. 10 highest correlations between Coh-Metrix variables and group A

<table>
<thead>
<tr>
<th>Variable</th>
<th>( r ) value</th>
<th>( p ) value</th>
<th>( r ) value</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSA given/new, sentences, mean</td>
<td>-0.486</td>
<td>&lt; 0.010</td>
<td>-0.370</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>Lexical diversity, MTLD, all words</td>
<td>0.444</td>
<td>&lt; 0.010</td>
<td>0.593</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Text Easability PC Temporality, z score</td>
<td>-0.429</td>
<td>&lt; 0.010</td>
<td>-0.127</td>
<td>&gt; 0.400</td>
</tr>
<tr>
<td>Temporal cohesion, tense and aspect repetition, mean</td>
<td>-0.424</td>
<td>&lt; 0.010</td>
<td>-0.150</td>
<td>&gt; 0.300</td>
</tr>
<tr>
<td>Meaningfulness, Colorado norms, content words, mean</td>
<td>-0.417</td>
<td>&lt; 0.010</td>
<td>-0.358</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>LSA overlap, all sentences in paragraph, mean</td>
<td>-0.407</td>
<td>&lt; 0.010</td>
<td>-0.315</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>LSA overlap, adjacent sentences, mean</td>
<td>-0.387</td>
<td>&lt; 0.050</td>
<td>-0.224</td>
<td>&gt; 0.100</td>
</tr>
<tr>
<td>Minimal Edit Distance, all words</td>
<td>0.373</td>
<td>&lt; 0.050</td>
<td>0.226</td>
<td>&gt; 0.100</td>
</tr>
<tr>
<td>Coh-Metrix L2 Readability</td>
<td>-0.343</td>
<td>&lt; 0.050</td>
<td>-0.524</td>
<td>&lt; 0.010</td>
</tr>
<tr>
<td>CELEX Log frequency for all words, mean</td>
<td>-0.333</td>
<td>&lt; 0.050</td>
<td>-0.547</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Temporality is an index of text easability, measuring cues and consistency of a text’s tense and aspect use. A high temporality value makes a text easier to process, and increases a reader’s situation model understanding (McNamara et al., 2014). Group A had a significant negative correlation of -0.429 between temporality and grades. This is yet another measure of explicit cohesive devices that is negatively associated with grades. Group B on the other hand had no clear correlation for this index. A similar index is found in the situation model bank of variables: temporal cohesion. The situation model is mentioned in the method section, but essentially “comprises the reader’s mental representation of the deeper underlying meaning of the text” (McNamara et al., 2014, p. 68). Temporal cohesion is important for temporal understanding in this mental representation. As is clear in Table 3 both temporality indices have very similar r values, for group A, and no significant results were found in group B. That being said, the average temporal cohesion score for both groups was almost identical: 0.794 in A and 0.786 in B and the standard deviation was also similar. There was in other words no clear difference in actual use of temporal cohesion between the two student groups; what differed was the correlation with grades. One possible reason for this is that teacher A simply sees temporality as more important when grading.

Coh-Metrix L2 Readability showed significant correlations for both groups, even though it was not represented in Table 3. In group A it was negatively connected to grades with an r value of -0.343, and in group B with the strong correlation score -0.524. L2 readability is a formula that attempts to predict readability for second language readers. It considers syntactic similarity, word frequency, word overlap and cohesion between sentences (McNamara et al., 2014). In a study of news texts simplified for English learners, the formula distinguished between three text levels (advanced, intermediate and beginning) at a success rate of 59% (Crossley, Allen & McNamara, 2011a). This index can then to some degree determine the general proficiency level of a text. It is therefore not at all surprising that high readability scores correlated with low grades, just as it successfully predicted easy to read beginner’s texts in Crossley et al. (2011a). The beginner texts in Crossley et al. (2011a) had a mean L2 readability score of 19.951, the intermediate 16.076 and advanced 12.897. The texts in the current study had a mean of 22.026 (A) and 20.333 (B). This means that both groups had a higher readability index than easy to read texts aimed at beginning learners. If seen as a measure of advanced language, group B, who had a lower grade average, in fact used more advanced language. With that said, one should be careful of using this index in such a fashion, as it was designed to assess readability and not proficiency.

3.2.2 Lexical sophistication

The second highest correlation regards lexical diversity, but unlike the vocd index in Tables 2 and 4, the highest lexical diversity index for group A was MTLD. The two indices are quite similar in their function. Lexical diversity is measured by type (unique words) token (all words) ratio (TTR), but this correlates strongly with text length. MTLD and vocd use algorithms to avoid this, but in different ways. “MTLD is calculated as the mean length of sequential word strings in a text that maintain a given TTR value” (McNamara et al., 2014). Vocd has been explained in section 3.1. Still, they both indicate the same thing, viz. that word variation is more common in highly rated essays. While the
The fact that a lexical diversity index is second highest in Table 3, and highest in Table 4, clearly shows it is important regardless of group. Crossley and McNamara (2012) found a correlation of 0.427 for lexical diversity and McNamara et al. (2010a) found that MTLD was among the top three most predictive variables of essay rating. However, Westerlund (2019) found very contradictory results. The study of example texts found no significant results, but in the set of 30 student essays lexical diversity was significantly (p < 0.03), negatively correlated to grades by -0.384. This difference is difficult to explain. The study analysed essays by adults doing an upper secondary school course, so an interesting possibility is that adult students learn English differently than teenagers. This would be interesting to study further. The results in this study are nevertheless in line with most previous studies.

Two other lexical indices are presented in Table 3. The lowest correlating index CELEX log frequency has been discussed above. Meaningfulness on the other hand only appeared among the highest correlations for group A. It is a psychological index that measures the average meaningfulness of words. Highly meaningful words are more closely associated with other words (McNamara et al., 2014). Further, the index is an important indicator of word knowledge (Crossley et al., 2011b). In the student essays, usage of words with lower meaningfulness was more frequent in higher rated essays. In group A the correlation was -0.417 and in group B -0.358. Therefore, words with looser connections to other words were indicative of a high grade, which is further evidence that lexical sophistication is an important part of perceived quality of student writing.

3.2.3 Syntactic complexity

The only measure of syntactic complexity in either table is minimal edit distance, all words. It is a measure of semantic and syntactic dissimilarity, counting the distance between the same word in consecutive sentences. Low distance between words would mean the text uses a word in the same syntactical pattern. Although minimal edit distance, all words is a measure of syntactic complexity, it is also associated with referential and semantic cohesion (McNamara et al., 2014). The r value for group A was 0.373 (p < 0.050), but no significant results were found in group B. A high value of this index means a greater distance between similar words in consecutive sentences, so the moderate correlation in group A means that repetition of the same word in a similar syntactic position was more common in lower rated essays. Since, minimal edit distance, all words is often correlated with cohesion, this reinforces the view that explicit cohesive devices are negatively connected to grades. Further, this means that repetition of words necessitated a greater syntactical variation for higher rated essays in group A. Why this was not the case in group B is difficult to say. It should however be noted that group B did have a significant correlation with another syntactic complexity index. Number of modifiers per noun phrase had an r value of 0.348. It is not represented in any of the tables because of its relatively low correlation score, but it shows that syntactic complexity had some connection to grades in group B as well. This indicates that while no strong connections were found between essay grade and syntactic complexity, it nevertheless had some role.
3.3 Group B

In Table 4 the ten highest correlations for group B in isolation are presented. Six of these have already been discussed in section 3.1. Since correlation scores were generally higher in group B than A, group B of course had a bigger impact on the correlation results for all essays as one group. One index, L2 readability, was discussed in section 3.2. Of the remaining three indices one is very similar to word count in Table 2, namely sentence count. It is essentially a shorthand for essay length and in group B it was the second most important variable to grade correlation. Just as with word count in table 2 no significant results were found for group A. This clear difference in the importance of essay length is plausibly explained by the relative uniformity of length in group A, especially compared with the much bigger standard deviation in group B, as discussed in section 3.1.

**CELEX word frequency for content words** compares all words in a text with their frequency in the CELEX database of 17.9 million words. The logarithmic frequency that was discussed in section 3.1 better reflects the lexical impact of reading speed while this index contains the raw data (McNamara et al., 2014). The strong correlation of -0.591 shows that use of less common words was more frequent in higher rated essays, and thus reinforces the idea that lexical sophistication is important for graders. However, no significant result was found in group A. If we add log frequency to this however, significant correlations were found in both groups, so while use of infrequent words was clearly more important in group B, words that affected reading speed had an impact on the grading of group A as well. Crossley and McNamara (2011) found a correlation value of -0.336 for CELEX content word frequency which is quite a bit lower than group B in the current study. Westerlund’s (2019) student essays were in between at -0.469 but the analysed example texts had an even higher correlation score of -0.688. These rather varied results show that word frequency seems to be important to perceived quality, but just how important varies in different groups.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group B</th>
<th></th>
<th>Group A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical diversity, VOCD, all words</td>
<td>0.622</td>
<td>&lt; 0.001</td>
<td>0.426</td>
<td>&lt; 0.010</td>
</tr>
<tr>
<td>Sentence count, number of sentences</td>
<td>0.618</td>
<td>&lt; 0.001</td>
<td>0.006</td>
<td>&gt; 0.900</td>
</tr>
<tr>
<td>CELEX word frequency for content words, mean</td>
<td>-0.591</td>
<td>&lt; 0.001</td>
<td>-0.235</td>
<td>&gt; 0.100</td>
</tr>
<tr>
<td>Content word overlap, all sentences, proportional, mean</td>
<td>-0.581</td>
<td>&lt; 0.001</td>
<td>-0.275</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>Familiarity for content words, mean</td>
<td>-0.577</td>
<td>&lt; 0.001</td>
<td>-0.208</td>
<td>&gt; 0.100</td>
</tr>
<tr>
<td>Text Easability PC Referential cohesion, z score</td>
<td>-0.560</td>
<td>&lt; 0.001</td>
<td>-0.328</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>CELEX Log frequency for all words, mean</td>
<td>-0.547</td>
<td>&lt; 0.001</td>
<td>-0.333</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>Coh-Metrix L2 Readability</td>
<td>-0.524</td>
<td>&lt; 0.001</td>
<td>-0.343</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>Hypernymy for nouns, mean</td>
<td>0.523</td>
<td>&lt; 0.001</td>
<td>0.278</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>Text Easability PC Narrativity, z score</td>
<td>-0.512</td>
<td>&lt; 0.001</td>
<td>-0.170</td>
<td>&gt; 0.200</td>
</tr>
</tbody>
</table>
The final index that has not been mentioned previously is *familiarity*. It measures how familiar words are perceived to be by adults. Sentences with high familiarity scores are processed more quickly (McNamara et al., 2014). This index is consequently similar to *CELEX word frequency*, but is a psychological measure tested on people rather than frequency in a database. Nevertheless, the results were quite similar to *word frequency*. Group B had a negative correlation of -0.577 while no significant correlation was found in group A. Again, Crossley and McNamara (2012) had a lower but significant score of -0.400 and Westerlund (2019) found a correlation of -0.379 in the student texts and the very high -0.700 for the example texts. This shows one more time the importance of lexical choice and sophistication to grading. Interestingly, the grading of group A was not as influenced by frequency and familiarity, but the second highest correlation for the group was still lexical but concerned lexical diversity (see table 3). The fact that *CELEX log frequency* correlated with grades for group A as well does, however, suggest that the use of less common words had an impact on grading.

### 3.4 Concluding comparison of group A and B

When comparing the independent results for group A and B there are some striking differences. Six of the ten strongest correlations for group B showed no significant results in group A. To a lesser degree the reverse is also true, as no significant results for three of the ten highest correlation scores for group A were found in group B. Only two indices were among the top ten correlations in both groups, although both groups also had very similar lexical diversity indices. A clear difference between the groups was the strength of correlations. The ten highest correlations for teacher A were between 0.343–0.486 while teacher B was between 0.512–0.622. There are some similarities and differences when looking at what type of indices were strongest for the respective groups. Five indices in group A had a clear connection to explicit cohesive devices, but only two in group B. Group B had more indices connected with word choice instead, five compared with three in A. That being said, word frequency and familiarity does not only reflect lexical sophistication but cohesion as well, so the difference was not as big as one might think. Only group A had a significant correlation between grades and a syntactic complexity index, while length mattered more for group B. Both groups, however, had a significant correlation with *L2 readability*. In sum, while the groups had some clear differences, not least in correlation strength, cohesion and lexical sophistication were nevertheless the most common distinguishers of essay grade level for both teachers’ groups.

### 4. Discussion and conclusion

#### 4.1 The implications of this study for the Swedish school

This study has shown that lexical sophistication correlates clearly with essay grades in two different groups of students with two different teachers, while increased use of explicit cohesive devices correlates negatively with grades. The relationship between these results and the knowledge requirements teachers in Sweden are supposed to base
their assessments on will now be discussed. Detail explanations for knowledge requirements for English 6 are provided by Skolverket (English). In addition to these explanations, the assessment guidelines for the test include factors for assessment. The linguistic factors to consider are: communicative strategies; fluency and free expression; adaptation to purpose, recipient, situation and genre; diversity, variation, complexity, clarity and certainty. For the final point, the following items should be considered: vocabulary, phraseology and idiomatic language; syntax, cohesion and structure; grammatical structures; spelling and punctuation (Skolverket, Skriftlig produktion och interaktion, Engelska 6). Both cohesion and lexical sophistication are mentioned in this guide. Diversity and variation in the guide can both be connected with cohesion and word choice as well. *L2 readability* might have a connection with fluency and free expression and *narrativity* is connected to adaptation to purpose. The list does not, however, offer advice on how to assign grades. For grade assignment the knowledge requirements are needed, but they are not as specific as the assessment guideline. There is a clear progression from level to level, such as for E writing should be “relatively varied”, simply “varied” for C and “varied [and] balanced” for an A (Skolverket, English). What this means is to a large extent up to the professional discretion of teachers. While the knowledge requirements of course influence how language teachers assess the work of their students, their vagueness probably leads to different types of assessments by teachers across the country. That is what the clear differences between the two groups in this study indicate. Gustafsson et al. (2014) have also shown that grading reliability is low for essays.

This study has focused on how two different teachers approach writing assessment, rather than how reliable their grading is. Therefore, not very much can be said of grading reliability for the two groups, however, some suggestions can be extrapolated from the Coh-Metrix results. As has been mentioned, *L2 readability* is based on a formula collecting different factors important to readability for language learners. The result from that index suggests that group B in fact used more advanced language than Group A, even though B’s average grade was about one point lower than A. Together with Brimi’s (2011) results that showed that grading of the same essay varied on average, what is equivalent to about one grade, it does not seem unlikely that the difference between the two groups was to some extent due to the different teachers, rather than the different students. Again, a study of reliability would be necessary to confirm this.

What this study has shown is that Coh-Metrix can identify how two different teachers in Sweden do value similar linguistic variables when assessing writing, which is also similar to international studies. To address the problem of reliability identified by Gustafsson et al. (2014), and possibly implied in this study, it is the contention of this paper that an automated essay scoring system (AES) should be developed for the Swedish school to be used as consultation in assessment for teachers. This study has shown the power of natural language processors, and previous studies have shown the reliability of AES (Landauer et al., 2003; Cohen et al., 2018). For a good overview of the validity, reliability and limitations of AES see Attali (2013). Wilson et al. (2017) have even demonstrated the possibilities of AES for formative assessment. It should be noted that the general view of AES is that it should be used as a complement to the more holistic evaluation of human assessors. There would undoubtedly be challenges in implementation of AES, especially
since there is a lack of research on the subject in Sweden. Instruction must, for instance, not be directed at pleasing the AES system, which is one reason for why a licensed teacher should have final authority over grading. Implementation would require further research, but it is nevertheless the contention of this study that to increase grading reliability across Sweden, and thus fairness for students—for whom upper secondary school grades have a big impact on the future—schools need to embrace the possibilities of AES. Especially as the national test is moving towards being completely digital, and as teachers often have a high workload.

4.2 Implications for field of research

The results of the current study will now be situated in the field of previous research. The importance of linguistic sophistication to perceived quality has been demonstrated both for L1 and L2 writers (McNamara et al., 2010a; Crossley & McNamara, 2012), but more recent studies have had different results. Perin and Lauterbach (2018) found no significant relationship for low level college writers and MacArthur et al. (2019) showed a negative correlation between grades and syntactic complexity for basic college writers. The proficiency difference between the students in the different studies could be a possible explanation for the disparate results. This would be consistent with the current study, in which the students too are below the writing proficiency level of ordinary college students. However, as discussed in section 3.2 some positive correlation for syntactic complexity was found in both groups, although the effect was small. This study in no way clarifies the role of linguistic sophistication to perceived essay quality, but further demonstrates that it seems to vary in different student groups and proficiency levels.

What is less complicated is lexical sophistication. It has consistently been shown to have a big impact on perceived essay quality (McNamara et al., 2010a; Crossley & McNamara, 2012; Perin & Lauterbach, 2018; MacArthur et al., 2019; Westerlund, 2019). Two of the three most predictive indices in McNamara et al. (2010a) were lexical: lexical diversity and word frequency. The same was true for group B in the current study, as was lexical diversity in group A.

A more complex factor is cohesion. This study found a very clear negative correlation between explicit cohesive devices and essay grade, which is in line with some previous studies (Crossley & McNamara, 2012; Perin & Lauterbach, 2018; Westerlund 2019). Crossley and McNamara (2012) found a negative correlation between grades and cohesion, and suggested that it could be explained by “a reverse cohesion effect” (p. 130). They state that low skilled readers benefit from explicit cohesive devices, but that the reverse is true for advanced readers. As advanced readers, English teachers would consequently prefer low cohesion texts, even though it is not preferable for beginning readers. Further, genre is a factor that should be considered. A study of two different essay types found different significant indexes depending on genre. In persuasive essays, the referential cohesion index argument overlap in all sentences could discriminate between low and high quality. In summaries, this was done by content word overlap, adjacent sentences, also a measure of referential cohesion (Perin & Lauterbach, 2018). Some form of cohesion was nevertheless significant in both genres. The current study was made on argumentative or discussion essays, and no group showed a strong correlation with the
index that mattered for persuasive essays. Instead, *content word overlap* which was significant for summaries had a correlation of -0.581 in group B and -0.434 for all essays. There is no clear pattern for the effect of different variables depending on genre, so this is an interesting topic for further study.

There are also studies with different results concerning cohesion. McNamara et al. (2010a) found no correlation and MacArthur et al. (2019) saw a positive connection. Crossley et al. (2016) concluded that “coherence for expert raters is a property of global cohesion and not of local cohesion” and specifically global cohesion correlated positively with grades in their study. However, the negative correlation in Perin and Lauterbach (2018) was also specifically based on global cohesion. The current study has not discriminated between the two types of cohesion and can therefore not illuminate this point. It is difficult to explain these differences other than to say that different groups seem to behave differently, so the effect of cohesion on perceived quality is an interesting topic for further investigation.

**4.3 Conclusion**

This study asked what language variables correlate with grades and how two teachers differ in their value of quality. Lexical sophistication clearly showed significant correlation with grades. Lexical diversity had the strongest lexical correlation in both groups. *MTLD* had an r value of 0.444 in group A, while *vocd* had an r value of 0.622 in group B and 0.428 in all essays. Word frequency was also impactful. *CELEX word frequency* had a -0.591 correlation in group B and *CELEX log frequency* had a lower but significant r value of -0.333 in group A, and -0.380 for all essays. Hence, usage of less common words was more common in higher quality essays. The other clearly important factor was explicit cohesive devices, but different indices demonstrated this in the two groups. Group A showed a negative correlation to grades with LSA and temporality, which group B did not. Referential cohesion was more important in group B, demonstrated by *content word overlap* (-0.581) and *text easability PC referential cohesion* (-0.560). So, explicit cohesive devices were negatively connected to grades in this study, but in different ways. This further illustrates the point in the previous subsection that groups differ in the importance of cohesion. Somewhat surprisingly, syntactic complexity was not an important factor. Contrary to McNamara et al. (2010a), *number of words before the main verb* did not correlate with grades. However, in group A *minimal edit distance, all words* correlated (0.373) with grades and in group B *number of modifiers per noun phrase* showed a 0.348 correlation. This means that syntactic complexity did, after all, have some connection to grades, but not as strong as other variables. A very clear difference between the two teachers was strength of correlations. The correlations in group A in Table 3 were between 0.343–0.486 while Table 4 showed correlations for group B between 0.512–0.622. The effectiveness of using Coh-Metrix to analyse grading patterns thus differs between groups. An important reason for this difference is of course that group A had much more uniform grading. In sum, the strongest correlations were mostly connected to explicit cohesive devices or word choice in both groups, although the indices differed. So, while teachers seem to differ in their evaluative focus, they do in fact value similar aspects of student writing.
References


