“SHE IS SUCH A B!” – “REALLY? HOW CAN YOU TELL?”

A qualitative study into inter-rater reliability in grading EFL writing in a Swedish upper-secondary school

JOSEFIN MÅRD GRINDE

School of Education, Culture and Communication
Degree project ENA308 15 hp
Supervisor: Olcay Sert
Examiner: Thorsten Schröter
Autumn 2018
Abstract

**Keywords:** English writing assessment, Sweden, upper-secondary school, inter-rater reliability, rater factors, rubric measurement, consistency in rater assessment in an ESL environment.

This project investigates the extent to which EFL teachers’ assessment practices of two students’ written texts differ in a Swedish upper-secondary school. It also seeks to understand the factors influencing the teachers regarding inter-rater reliability in their assessment and marking process. The results show inconsistencies in the summative grades given by the raters; these inconsistencies include what the raters deem important in the rubric; however, the actual assessment process was very similar for different raters. Based on the themes found in the content analysis regarding what perceived factors affected the raters, the results showed that peer-assessment, assessment training, context, and time were of importance to the raters. Emerging themes indicate that the interpretation of rubrics, which should actually matter the most when it comes to assessment, causes inconsistencies in summative marking, regardless of the use of the same rubrics, criteria and instructions by the raters. The results suggest a need for peer-assessment as a tool in the assessment and marking of students’ texts to ensure inter-rater reliability, which would mean that more time needs to be allocated to grading.
# Table of Content

1 Introduction ........................................................................................................... 1
   1.1 Purpose ........................................................................................................... 2

2 Background ............................................................................................................ 3
   2.1 Definition of terminology .............................................................................. 3
   2.2 The Swedish curriculum ............................................................................... 3
   2.3 Previous research ............................................................................................ 4
      2.3.1 The history of writing assessment ......................................................... 4
      2.3.2 Factors that might affect raters .............................................................. 4
      2.3.3 English writing and assessment .............................................................. 5

3 Method .................................................................................................................... 6
   3.1 Data collection ................................................................................................ 6
      3.1.1 Written text ............................................................................................. 7
      3.1.2 Open-ended survey ................................................................................. 7
      3.1.3 Focus group ............................................................................................. 7
      3.1.4 Challenges ............................................................................................... 8
   3.2 Data analysis .................................................................................................... 8
   3.3 Ethical principles ............................................................................................. 8

4 Results .................................................................................................................... 9
   4.1 Rater assessment process and grades ............................................................ 9
   4.2 Professional and educational factors ............................................................. 14

5 Discussion and Conclusion .................................................................................. 16

References .................................................................................................................. 20

Appendix A: Student text A ..................................................................................... 22
Appendix B: Student text B ..................................................................................... 23
Appendix C: The assignment ..................................................................................... 25
1 Introduction

When I was teaching a class last semester in Swedish, the students had an ongoing discussion regarding what grade they might get, how to improve the received grade and at times almost seemed to be competing with one another. When the final grade was posted and the conversation with the students took place, some of the girls came in as a group. One of them got the grade B. Her friends all started laughing and said she was such a B. This was apparently her highest grade. My colleagues and I could not help but laughed, due to the meaning of someone being a B, as in ‘bitch’. This, however, started a long discussion among the group of teachers in the room regarding grades and the perceived difficulties in inter-rater reliability when grading one’s students.

The grading system in Swedish schools today is under constant scrutiny by students, parents, the teachers and the National Agency for Education (Skolverket). There is a constant discussion regarding what is being graded and what a grade stands for. Quinn (2013) argues that “a grade is a piece of information that attempts to report something about a student’s education” (p.5). Looking at this very general definition, it is easy to see how teachers at times may find it difficult to interpret what to do with the information they have about a student’s education and how to set consistent and fair grades. In this case, consistent means that a grade will be the same, regardless of the teacher who assesses the assignment whether the assignment is being peer-assessed, assessed by one rater, or assigned by one teacher and rated by another. Appropriate assessment or “good assessment of writing”, as it is called by Hamp-Lyons (2012), matters because “the written word remains a principal medium of communication and one of the hopes for achievement of understanding between people everywhere” (p.6).

To better achieve consistency, teachers, or raters as they may also be called when the focus is on assessment and marking, use different tools to help them. Raters use, among other things, rubrics and peer-assessment as well as written and oral feedback. This can be done in the classroom, using computer software, and together with other raters. Meetings can be held to ensure that every student will be assessed; the raters can seek help within the school as well as to educate themselves outside of their workplace. So how is it that researchers have found inconsistencies in raters’ reliability (Quintero & Rodriguez, 2013; Kayapinar, 2014)? Hamp-Lyons (2002) looks into some of the outside factors that might affect raters, such as the
political situation and technological advancements. This study, however, will look into the educational and professional factors, which include assessment training and the use of rubrics that might cause an inconsistency in assessment.

1.1 Purpose

When grading and assessing students of English as a foreign language (EFL), raters’ behavior can be affected by educational, professional, linguistic, and personal factors (Quintero & Rodriguez, 2013). Upper-secondary students of English in Sweden are assessed and graded following the guidelines of the curriculum called *Läroplan för gymnasieskolan 2011, examensmål och gymnasiegemensamma ämnen*, known as Lgy11 (2010). Against this background, the aim of this study is to gain a better understanding of the perceived professional and educational factors that might affect inter-rater reliability as well as the grading process of the participating raters. Nine raters were involved in this study, each taking part in sections of the study that they were able to, depending on their workload at the time.

To gain a better understanding of these factors and the grading process, two texts written by students in an EFL environment were given, together with the assignment and the rubric for assessment, to four raters who provided a written summary of their thinking process as well as a summative grade for each text. In addition to this, five raters were given an open-ended survey regarding their assessment practices and professional and educational background.

Seven raters then attended a focus group, two from the original group and five new raters, and discussed the assignment, the rubrics, and the texts regarding their focus while rating, reflecting on factors that might affect their processes and grades given. This study, therefore, focuses on the assessment process of the raters, as well as possible inconsistencies, or consistencies, in the raters’ summative grading.

Based on this purpose, the following research questions are posed:

- To what extent do the EFL teachers’ assessment and marking processes of two students’ written texts differ?

- What are the perceptions of the EFL teachers regarding the factors that affect inter-rater reliability when assessing and marking students’ written texts?
2 Background

2.1 Definition of terminology

This study examines the factors that might affect consistency, or inconsistency, in raters’ grading of EFL writing. The term *intra-rater reliability*, according to Quintero and Rodriguez (2013), addresses a single rater’s consistency, unlike *inter-rater reliability*, which refers to how one rater varies from another rater in consistency (p.65).

2.2 The Swedish curriculum

As pointed out above, Swedish teachers in upper-secondary schools follow a curriculum often referred to as Lgy11 (Skolverket, 2010) when assessing students, whether it is through written or oral assignments. One of the goals of Lgy11 (Skolverket, 2010) is related to formative assessment of students throughout a course. This is done to help raters adapt their teaching to best suit the students’ needs in order to progress in their educational achievements. Lgy11 specifies the criteria for each grade, depending on subject. Skolverket also publishes recommendations regarding teachers’ work, to help further equality in grading. Recently, a new set of recommendations was released (Skolverket, 2018) that addressed grading and assessment aspects of Lgy11 and how teachers can interpret the criteria. Both formative and summative assessment and grading are addressed.

Skolverket (2018) advocates formative assessment by teachers to ensure validity and reliability in any assessment situation. Formative assessment is to be used to help students’ progress over time, as part of a course or program. As for summative assessment and grading, Skolverket (2018) notes that there has to be enough variation in assessment situations, which could be achieved by providing different assignments to the students. The variation can increase reliability in raters’ assessment, and provide validity to any grade so it will not be up for questioning. This can also be done, according to Skolverket (2018), by the rater using documentation and assessment based on the rubrics in Lgy11 (Skolverket, 2010). As for the summative grading of individual assignments, as in connection with the assignment used in this study, it is not something that Skolverket (2018) recommends, due to the fact that it would not be valid at the time of grading for the completion of a course when thinking from a progression and formative standpoint.
2.3 Previous research

Inter-rater reliability has been addressed in many studies, and previous research in the field includes aspects as diverse as factors that affect university professors’ reliability when grading English as a Second Language (ESL) graduate students (Huang & Foote, 2010), the validity of scores from raters grading native speaking and EFL students (Huang, 2012), and non-expert raters’ accuracy in scoring an essay based on content and possible differences compared to expert raters (Wolfe, Song & Jiao, 2016). Only few studies, however, have focused on upper-secondary school students in an EFL environment and the factors that affect reliability in EFL raters when grading and assessing written assignments of EFL students.

2.3.1 The history of writing assessment

The history of using writing as a way of furthering knowledge and education, according to Hamp-Lyons (2002), could be traced back to Christian Europe and the monasteries, with a focus on Greek and Latin, from the Middle Ages to the present date. Oral exams gave way to written exams as the most common practice of assessment at universities such as Oxford and Cambridge (Hamp-Lyons, 2002). In the United States of America, as Hamp-Lyons reports, Harvard led the way to the inclusion of written exams, thus replacing the previous oral admission tests. Written assignments in schools need to be assessed by raters, and that assessment can be inconsistent, due to different factors.

2.3.2 Factors that might affect raters

Quinn (2013) tries not only to define the inner workings of grades in general but also discusses problems that raters might face when evaluating students. One of those problems is the difference between students’ actual skills and knowledge and their performance in e.g. a test. When connecting this to inter-rater reliability, Quinn (2013) writes that

aiming for consistency of grading at any level is a very tall order given the seeming impossibility of getting so many teachers to apply to the same standard – standard that are often subjective – to their assessment of student work. (p. 83)

Zhang (2016) goes as far as to write that “rater variability has long been regarded as one of the most significant sources of measurement error and a potential threat to the reliability and fairness of performance assessment” (p. 37). Zhang also found that accurate ratings could be affected by the use of raters’ cognitive and meta-cognitive strategies in rater assessment training. e.g. interpretation strategies and judgment strategies. Zhang (2016) uses a descriptive
approach when looking at the meta-cognitive processes of self-monitoring focus, and the
cognitive processes of rhetorical and ideational focus as well as language focus.

Previous research has shown that there are inconsistencies in raters’ assessment and grading
of English writing (Quintero & Rodriguez, 2013; Kayapinar, 2014), but that improvement can
be achieved regarding some of the factors that affect raters (Hamp-Lyons, 2002). Some
aspects that might affect writing assessment, according to Hamp-Lyons (2002), include
“technological, humanistic, political, and ethical concerns” (p.10). The author examines
factors such as technological advancement that are considered outside of the rater and that
might affect the reliability, concluding that raters need to keep up with the digital age without
losing sight of the fact that it is humans being rated. Other factors that have been found to
affect inter-rater reliability concern educational background, differences in the raters’
linguistic background, whether the rater is a native or non-native speaker of English, as well
as previous training in assessment (Quintero & Rodriguez, 2013). One other factor that has
been addressed by Hamp-Lyons (2002) is the political situation, which affects education.
Based on political standpoints, and a desire to enhance political power based on test scores
and numbers, decisions can be made that may not be fruitful for the educational system,

Assessments are subjective and filled with value-based perceptions from the rater based on
professional, personal and outside factors (Quintero & Rodriguez, 2013; Hamp-Lyons, 2002;
Kayapinar, 2014). The goal in assessing students, according to Quintero and Rodriguez
(2013), is to “reflect students’ abilities rather than unrelated factors” (p.1). This is done with
the help of different types of rubrics and even though these assessment tools have been
created to help raters stay objective and to promote accuracy and validity, there are still
inconsistencies in inter-rater reliability. The grading process of raters and some of the
educational and professional factors that affect that process will be examined in this study
with the help of an open-ended survey, written feedback on writing assignments from raters,
and a focus group.

2.3.3 English writing and assessment

A study conducted by Johnson and Van Brackle (2012) examines raters’ perception of errors
made by African American English writers in comparison to Standard American writers and
ESL writers using holistic assessment. The study concluded that raters were more lenient
towards errors made by ESL writers than African American writers, due to the fact that
language errors were somewhat expected by ESL students, but not by African American writers (p. 21). Barkaoui (2012) examined whether or not evaluation criteria would change with the raters’ experience and found that novice raters were more lenient than experienced raters and that experienced raters “referred more to evaluation criteria other than those listed in the rating scale” (p.1). Both the experienced raters as well as the novice ones seemed to focus on the commutative aspects of writing although the argumentative parts where of bigger importance to the novice raters than the more experienced raters. In another study, Kayapınar (2014) argued that ensuring rater reliability in English writing could be done by holistic rating, the use of an essay assessment scale, essay criteria checklists or by using rubrics similar to those employed in the present study. A study that showed acceptable results in inter-rater reliability in English writing assessment was Saxton, Belanger and Becker (2012), where two raters blindly scored students’ work and even though one rater needed more assessment training, the intra-rater reliability showed consistency (p.1).

3 Method

This section will focus on the procedures of conducting this study, the data collection, and the ethical principles applied. The data were collected at an upper-secondary school in Sweden, using an open-ended survey, written feedback on writing assignments from raters, and a focus group. The data was then analyzed using content analysis based on Mackey and Grass (2012).

3.1 Data collection

A class of students were given an assignment by their teacher, namely to write an informal letter about the movie Red Dog (2011). Two of these letters were picked out (see Appendix A and B respectively), based on the difference in the grade they received, and used in this study. Ten raters were asked to participate in this part of the study, of whom seven agreed and four eventually completed a written response within the deadline.

As a second step of collecting data for this study, an open-ended survey was given to participants. Ten raters were asked to participate in the survey and five of them completed the assigned survey within the given time. Out of those five participants, four submitted a written response to the students’ texts within the deadline and included a summative grade.

Lastly a focus group was held with seven English teachers, who were presented with the written texts of the students, the assignment and the rubrics (see Appendix C).
3.1.1 Written text

The teachers were given two different texts written by two different students in the students’ English class. The texts were graded by 4 separate raters for this study.

The material presented to the raters included the texts themselves (see Appendix A and B respectively), as well as the assignment (see Appendix C), which included standard guidelines from Lgy11 (Skolverket, 2010). The guidelines were rubrics, which covered five criteria for English writing. The raters were also asked, while grading, to fill out the rubric or include it in their full-length text.

The grading was done on an A-F scale and based on the criteria in Lgy11 (Skolverket, 2010). The raters were asked to submit a written response as to why they believed the texts deserved the grade that they had given them and explain the grading process. This submission was expected to be no longer than one page per text, per rater.

3.1.2 Open-ended survey

All of the raters that finished the written response and attended the focus group were asked to answer a 20-question open-ended survey. The questions were about possible factors that may have affected the grades given for this study, and their assessment process. These questions were based on previous research (Quintero & Rodriguez, 2013: Hamp-Lyons, 2002) and were concerned with the educational background of the teachers, professional background and previous training in assessment.

3.1.3 Focus group

The raters who were asked to participate in this study were also invited to attend a semi-structured focus group interview. The invitation to participate went out to all English teachers at the school in question. That meeting included a discussion of the assignment, the texts assessed, rater factors, assessment training, and possible inconsistencies in grading. The discussion, which took place in December 2018, comprised of seven participants apart from myself and lasted for 24 minutes. The participants had given their written consent for me to audio-record the meeting.
3.1.4 Challenges

Throughout this study, there were some issues regarding the number of informants that participated. Out of the ten raters who were asked to participate in the written response section of this study, only four actually submitted their answers. The same issues can be found regarding the survey. As to the focus group interview, however, several informants were willing to participate regardless of the previously experienced time constraints, and five new informants attended. These five new informants were part of the original ten raters to be asked to be part of the study, but had declined participation in the survey and written feedback parts.

3.2 Data analysis

A content analysis was carried out, based on Stukát (2014), Silverman (2004) and Mackey and Gass (2012), on both the texts submitted by the raters as well as on the transcribed recording of the focus group discussion. For this study the recommendations made by Mackey and Gass (2012) were followed, notably that “it is very important to be able to account for how and why you chose the themes identified and for choosing the examples included in the written report. Are these highly representative, or just the most interesting or amusing?” (p.108). As part of the content analyses and based on the texts submitted by the raters, as well as guidelines from Stukat (2014), a search for themes was conducted and the themes were sorted into categories. The coding was done continuously throughout the 2-week time period when the texts came in.

The quotations used in this study were translated from Swedish to English where needed and transcribed here in a slightly modified fashion to ensure an easier read; however, the only aspects altered are words used as fillers, incomplete sentences and interrupted turn taking. Each rater was given a number, from 1 to 7, to ensure anonymity, and the student texts were given the letters A and B.

3.3 Ethical principles

For transparency, all participants were informed about this study in accordance with the guidelines of the Swedish Research Council’s ethical guidelines (Vetenskapsrådet, 2002). The identity of any and all participants would remain anonymous, as would the name of the school and any other identifying features within this study. This is why all the participants’ names have been replaced with a number and the students’ texts have been given a letter.
As a further aspect of transparency, all communication between the involved parties is available for possible peer review, as is any other documentation regarding this project where there is no chance of identifying participants.

4 Results

In the following section, the results will be presented and divided into categories based on the themes that emerged in the data analysis process. Each section is based on relevant literature and the research questions. Section 4.1 will address the rater assessment process and 4.2 will address factors that affect raters, including assessment training.

Each rater was given a number in the focus group based on where they sat, and each rater was given a number when they submitted their written responses, which they kept for the survey.

4.1 Rater assessment process and grades

<table>
<thead>
<tr>
<th>Main themes</th>
<th>Use of rubric</th>
<th>What to address in a text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-themes</td>
<td>Clarity and guidance</td>
<td>Grammar, understanding and fluency</td>
</tr>
</tbody>
</table>

One of the research questions of this project focuses on gaining a better understanding of the differences between the grading processes of the raters who participated in the study. To examine this, contextual themes in the written responses, the survey and the focus group discussions were analyzed. The themes that could be found when looking at the raters’ assessment process were the use of rubrics and the choice of what to address in a text, i.e. what was of importance in the rubrics and whether it was present in the text. The sub-themes for what to address in a text were grammar, understanding and fluency, which were mentioned by all the raters regarding the rubric, whether they were found in the texts or not. For this study, the raters were asked to use the rubrics when grading and assessing the texts, though they were not told what specifically they should focus on. As part of their participation in this study, the raters were asked to give a summative grade (as Figures 1 and 2 illustrate) to each of the assessed texts. The grades given by the raters varied, as can be seen in figures 1
and 2, and no rater gave the same grade to both texts’ that they graded, it was a mixed response from all of the raters.

Figure 1. Grades given for text A.

Figure 2. Grades given for text B.

When looking at what themes emerged from the raters’ written responses in regard to what they focused on, fluency and grammar stood out. These themes stood out regardless of them being assessed as being positive or negative aspects of the texts. One rater wrote regarding text B:

A big plus for the excellent division into paragraphs, to me it cancels out the effects of occasional language mistakes. I would grade the text an A for understanding”

Whereas regarding text A the same rater wrote that
There are too many mistakes in the sentence structure to reach a higher grade, the student has a very ‘spoken’ language.

That seems to indicate that there are things more important to the rater than grammar, as evident in the comments on text B; however, if the grammar is so poor that it affects fluency, it becomes more important to the rater, as evident in the comments on text A.

Another rater wrote the following regarding fluency and grammar in text B:

The student has copy-pasted almost a whole paragraph […], which indicates that he/she is not fully capable of expressing him/herself in a satisfying way.

The rater also commented on fluency in text B by stating that:

“The sentences are quite short, which affects the fluency a bit; the student also mixes past and present tense, which is a bit disturbing, but not too much […] some minor errors, like apostrophe genitive and no capital letter at the beginning of sentences”

However, for text A, the same rater wrote:

There is fluency (although some sentences are quite long).

This indicates that to this rater, fluency is of greater importance than grammatical mistakes, and that the latter do not greatly affect fluency in this case.

As for students’ understanding in relation to the criteria in the rubric, one of the raters wrote that:

The student show in his/her text that he/she has understood the movie and based on the movie can give well based examples e.g. can account for characters and plot [translated from Swedish]

The rater indicated that, for text A, the student had understood the movie, based on multiple examples in the student’s text. As for text B, the rater wrote the exact same thing, despite giving the texts different summative grades. The student who got the lower grade, however, did not include certain themes in their text (e.g. friendship). The rater wrote, regarding text B:

The student reproduces details, but also themes such as friendship, which shows a sense of nuances and comprehension.
This was not part of the rubric criteria or the assignment, so this would suggest that the rater added a criterion that the students were not aware of, and therefore could not adapt their submitted text to. One of the raters talked about expectations that were not mentioned in the criteria, saying that the students did not provide sufficient information:

> don’t get anything on how the encounter with the dog affects you; in text B there is more information there

This indicated that text B was considered to be better written due to the amount of information given by the student. However, another rater responded as follows:

> but the instructions don’t say that you have to, they are more like a guidance, they might help you

The second rater seemed inclined to base their opinion of the text on the criteria and instructions, rather than adding personal opinions to the assessment process. The assessment process also seemed connected to the raters’ opinions on the assignment and criteria themselves. The emerging themes that were found when the rubrics and assignment were discussed in the written responses, the survey and the focus group, were clarity and guidance. The teachers explained that clarity and guidance could be achieved by adding an interpretation of the rubric and reducing the length of the assignment as a whole. One rater stated in the focus group that:

> they [the students] would not have interpreted this [the rubric] clearly, they would need more guidance

Another rater added that:

> it is difficult to know how much guidance to give them; some students might see this [the assignment] as a checklist; as long as I have answered all of these questions then I have completed the assignment

This shows that the raters themselves interpret the assignment and the rubrics differently and focus on different aspects of grammar, fluency, criteria, and assignment.

As for student comprehension regarding the assignment, the rubrics and what the raters found important, one rater said:

> I think at least sometimes I try to show them examples of different levels of essays, so they can see for themselves the difference between at least an E and an A
Another rater added:

especially when you reach English 6, because it is very difficult sometimes to explain what you expect from them at a higher level, then it is good to have an example ‘well this essay did get an A or a B, a higher grade’

A third rater added:

I try to always explain the context so if you are writing this kind of text, if it is a letter or an article or whatever it is, I am looking for this, or this is what we have worked on or this is what I am going to assess later, so they know

These statements indicate that the raters usually try to include the students in the assessment process and to clarify to them what they would expect and what exactly the students would be graded on. One rater explained the positive effects of using rubrics in assessment and grading, as:

some students are very persistent, they want to know ‘why did I get this grade and not this one’, and then you back this up with, well, you see the assignment was about this and central innehåll [core content] relates to that, and the rubrics relate to that and it is easier to explain I think, why you have given a certain grade

That way, the rater and the student can, with visual help, see which criteria the student has reached and what is expected for each grade and assignment, connecting back to the theme of clarity and guidance.
4.2 Professional and educational factors

<table>
<thead>
<tr>
<th>Main theme</th>
<th>Peer-assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sub-themes</strong></td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Context</td>
</tr>
<tr>
<td></td>
<td>Practical assessment</td>
</tr>
</tbody>
</table>

The second of the research questions of this study focuses on gaining a better understanding of the participating raters’ perceptions of factors that affect the assessment and marking processes of the selected students’ written texts. To examine this, contextual themes were analyzed. The theme that emerged when looking at the perceived factors was peer-assessment, with the sub-themes of time and context, as well as practical assessment.

One rater said that:

I always feel more confident when I have done that [peer-assessed]

to which another rater added:

with the system that we have here, the best way would be to get a second opinion, that is the only way

One rater explained the peer-assessment situation at the school as follows:

I ask the ones [other teachers] sitting next to me and if I know that someone is having the same course as me, then I can email those teachers sometimes. But I would like to get more time to do it. This is nice, just sitting here like we do now

Another rater reinforced the statement in relation to time management, claiming that:

we never have time for it [peer-assessment]

Time was one of the frequently mentioned concepts by the participants and emerged as a sub-theme of the broader theme “peer-assessment”. The same could be said for the theme of context, brought up by the raters themselves, not as part of the suggested factors.
The context factor, as described by one participant, concerns the individual assessment based on a group behavior, and the judgment by raters. One rater said,

I think the context affects you because if you have really good students around you all the time, then you expect, I mean, nothing can be good enough and even if something is on an E level, you might think it’s like an F but it is still on the E level. Same thing if you have really poor students, then it is the same thing because then you might actually think that that one can actually write a text, that one should have an A, so I think it might be good to have students in varied spectra. I think that is beneficial to all the students and for all teaching.

This quote indicates an assumption made by the rater, namely that one student’s achievements will sometimes be assessed in relation to another student’s achievement, not the actual criteria based on Lgy 11 (Skolverket, 2010) or the recommendations made by Skolverket (2018). This idea regarding assessment and grading could be connected to assessment training, which some of the raters felt they were lacking. One of the questions the raters were asked in the survey was, Did you receive any training or education in assessment and grading while you were studying for your degree? One of raters answered that:

in English we had a 7,5 hp course where we assessed Nationella Prov [standardized tests], the writing part, in English 5. In Swedish it was much the same. During my training (VFU) I got to do this a great deal in collaboration with my tutor at the school where I had my placement.

The second rater answered:

Yes, practical training, mostly.

This indicated that the training received, if any, by the raters was practically oriented and that the peer-assessment situations that occurred during the monthly meetings can be seen as an extension of previous training, or as one rater expressed it, in connection to assessment training in English writing:

You need to update yourself all the time, conferences, forums, open discussions.

Two raters were more vocal and detailed in their written responses and the focus group interview, to the extent that they became the focus point regarding assessment training and emerging themes. When the data was analyzed, the themes that emerged were, again, peer-
assessment with the sub-theme of practical assessment training. The two main contributors talked about the use of peer-assessment as an assessment training tool and how they would utilize the monthly meetings that were held with all the English teachers at the school: “

We sometimes do it [peer-assess] during these meetings, we copy something, maybe it’s from a national test, and assess

The second rater continued:

We are usually in agreement, it is rarely that we have a big conflict about something, but we often disagree with Skolverket. They give us free examples with what is an E or a C or an A and we are always amazed by their [interrupted]

The fact that the rater stated that the group of English teachers that meets once a month often agreed on a grade cannot be supported by the results of this study. The clear inconsistencies found in the summative grades given by the raters provided a somewhat different picture.

5 Discussion and Conclusion

<table>
<thead>
<tr>
<th>Research question 1:</th>
<th>Research question 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>To what extent do EFL teachers’ assessment and marking processes of two students’ written texts differ?</td>
<td>What are the perceptions of the EFL teachers regarding the factors that affect inter-rater reliability when assessing and marking students’ written texts?</td>
</tr>
</tbody>
</table>

**Main theme**: the use of rubrics, the choice of what to address in a text

**Sub-themes**: Clarity and guidance. Grammar, understanding, and fluency

**Main theme**: peer-assessment and assessment training

**Sub-themes**: time, context and practical assessment training

Both of the research questions were examined in the same way, by the use of written responses from raters, a focus group and an open-ended survey and the search for themes and sub-themes. For the first research question the assessment processes of the raters and how they may differ was the focus. In the data thus collected, some themes emerged, namely the use of rubrics, the choice of what to address in a text, i.e. what was of importance in the
rubrics and was it present in the text. The sub-themes found were grammar, understanding, and fluency. The aim of the second research question was to gain a better understanding of the perceived factors that affect EFL teachers in their assessment and marking process of students’ written text. The themes identified in connection with this were peer-assessment and assessment training, with the sub-themes of time, context and practical assessment training.

The results of the first research question indicated that the use of rubrics by the participants was consistent; however, the interpretations of what was expected from the students were not. This may have affected the inter-rater reliability for the summative grade given by the raters to each text. The results indicate that inter-rater reliability is limited. The raters’ grades ranged from D to B for both texts, and that gap in consistency could be connected to the expectation the raters had on a finished text that were never disclosed to the students. These expectations could concern content not provided by a student, such as:

how the encounter with the dog affects you, in text B there is more information there

This got criticized by other raters for not being specified in the information given to the students. On the sub-themes of grammar and fluency, there were inconsistencies found regarding not only inter-rater but also to intra-rater reliability. The raters’ assessment indicated that one text, text B, should have received negative feedback on fluency, had the rater shown consistency. However, despite the statements on grammar, fluency was only slightly affected and not in connection to grammar, but rather length of sentences. This inconsistency can be related to factors that may affect the assessment process of raters, inter-rater reliability and summative grades.

Not all perceived factors that may affect raters’ reliability in assessment and grading can be addressed within the scope of this study. The factors that are addressed, however, are based on previous research and include educational and professional background (Quintero & Rodriguez, 2013) as well as the use of different measurement tools such as rubrics, to ensure inter-rater reliability (Barkaoui, 2012; Saxton, Belanger & Becker, 2012; Kayapinar, 2014).

The other research question of this study focuses on gaining a better understanding of the participating raters’ perceptions of factors that affect the assessment and marking processes of the selected students’ written texts. The raters themselves noted that context was a vital factor for them in the process of grading. The context of “poor” and “good” students, based on the achievements of the students as a group, which could affect the raters’ assessment of the individual, was one factor. The raters agreed that an important step towards consistency in
inter-rater reliability could be found in peer-assessment. Peer-assessment was conducted when the raters could find the time and was sometimes integrated in their meetings. The factor of time, however, was also what affected how often students’ texts could be assessed and by whom. The raters discussed, in the focus group, the use of peer-assessment for students as a way of seeing the process and understanding the rubrics.

In conclusion, it seems that the factors that affect a rater’s reliability differ from rater to rater. The results show that time is considered a large factor, as well as context, and assessment training. Furthermore, the result of this study shows that there are clear inconsistencies in inter-rater reliability in assessment and grading in English writing at the school where this study was conducted. There are areas that show consistency such as the assessment process, but this does not seem to affect the inconsistencies in summative grading by the raters. To further the inter-rater reliability of raters, suitable steps should be considered, such as giving the raters enough time to peer-assess written texts, as well as increase the assessment training available for raters, including practical assessment training. As for the use of rubrics as a tool for increased inter-rater reliability and the specific use of the rubric presented in Lgy 11 (Skolverket, 2010), there are indications that the raters might find them helpful despite the fact that they do not seem to increase inter-rater reliability. A natural continuation of this study would be to examine how specific factors affect raters and how the use of rubrics could be altered to increase inter-rater reliability.
References


Appendix A: Student text A

Dear friend, today I experienced something out of the ordinary. As you know I recently moved to Australia and I just met an extraordinary dog, he was sitting by the road, like he was waiting for someone. I figured I’d wait and see, maybe he needed help. Not longer than five minutes after he got there, I could see a red car driving towards him from the horizon. They were driving like maniacs, I got scared they’d hit him. But as he passed the dog, he stopped, in the middle of nowhere, he opened the door, and the next thing I see is the dog jumping into the car. The driver closes the door and then he drove off.

Once I arrived in town, I decided to go to the bar to get a drink, the dog was already there. I asked the bartender, he said his name was “Red” he had a reputation in town, and everyone knew who he was. Red was a dog for everyone, he’d listen to everyone, especially when you needed someone who listened.

They said he was found as a puppy, he had no master for years, but then John came by, he saved Red one evening, as the whole town were betting on how fast he could eat certain things, they wanted to feed him a live chicken, but John told them to stop. He got into a fight with one of the guys in the bar that night, but then Red started barking he made everyone in the room stop what they did, He’d never acted like that before, that’s how they connected and how John became Red’s master. From that day they road along in Johns bus.

A while later, John had to pick something up at the nearby gas station. And while he’s away a young woman named Nancy showed up, she wanted a ride on the bus, but once she got aboard all the seats were taken by the workers but they all moved in hopes that she would sit beside them. Nancy decided she wanted to sit by Red dog, so she tried commanding him to sit down on the floor, but red never listens to anyone besides his master.

She tells Red to stop acting up and to act like a gentleman, when she said that Red looks at her with a “scared” face and steps off the other seat. Nancy sat down beside Red and all the other workers that hade made space for Nancy went back to their original positions.

I’ve heard so much about this dog, and I hope I have the chance to learn more about him. And if I do I will surely tell you more about it!

Love (name of student)
Appendix B: Student text B

Dear Sara,

Today I experienced something out of the ordinary. I am in Australia and have already spent one of my four weeks here. The weather is incredible, and the food taste delicious. I am so happy to be here! A lot of incredible things happened the first week and I am so excited. I met a dog yesterday and he was amazing. I have never met anyone like him before. His name is Red Dog and he has no owner. I met the dog when I was driving to the beach. He was standing on the road. I stopped and opened the back door, he jumped in to the car. He was very social and out spoken. Everyone calls him Red Dog because of his fantastic red fur. He told me that he is walking thru Australia. During his journey he talked to a lot of people. He had the ability to bring people together. I travelled to western Australia and he decided to come with me.

I have a lot of fun with him and I would like to tell you about it. One day when I was sick he came to my hotel and he gave me medicine and food. He stayed with me all day and I felt very lucky to have him. Another day when my car stopped in the middle of the forest and I could not start it. I did not know what I should do or where do I go. It was dark night and I was so scared. Fortunately, I had the red dog and he helped me find the way and he guarded me all the way to the hotel. That night went well.

After a week I met a person in the restaurant and he told the true story about Red dog. He said, “red dog belonged to everybody but to no one”. He said also that the dog was known simply as Red Dog. He was known for stopping cars on the road by walking right in the path of an oncoming vehicle until it stopped and then he would hop in and travel to wherever the car driver was going. He took bus rides as well and, once, when a new driver pushed him off her bus, the passengers all disembarked in protest.

My travel in western Australia was fantastic and I learn a lot of things about the Red Dog. I do not want to leave Red Dog but unfortunately, he left me. He died when he found out that White Dog, his big love had been murdered.

White dog was kind. She had fur, white as snow, and clear green eyes. Every dog in the neighbourhood wanted her, but she wanted no one else than Red Dog. The became a couple years ago and they were always together. They travelled, ate and slept. The first time they were apart was when the unthinkable happened. White Dog was walking down the street,
when a man shot her. Her fur was not only amazing but also unique, which meant constantly danger. White dog was gone, and her soulmate Red Dog could not accept her faith. He travelled round the world, trying to find her, but without success.

When he realized that he was never going to see her again, he collapsed. He became very sick and passed away. It was very hard for me to hear about my friend, Red Dogs, life. My meeting with him mean a lot to me. He taught me about courage and friendship. I wish that he was still alive.

Sara, I hope everything is alright with you.

Best regards, your (name of student).
Appendix C: The assignment

Red Dog

We have just seen the movie Red dog together in class, a movie that is set in Australia. Imagine now that you have moved to Australia and just met the famous Red Dog whilst he was out on his legendary journey across the country. Your task is to start writing a letter (1 A4) to a friend back in Sweden, telling him or her about your encounter with the dog. Relate to the movie and use your imagination!

Begin the letter with "Dear friend, today I experienced something out of the ordinary". The following questions might help you:

What did the dog do?
What was he like?
When and where did you meet him?
How did the encounter with the dog affect you?

Centralt innehåll

Kommunikationens innehåll

• Ämnesområden med anknytning till elevernas utbildning samt samhälls- och arbetsliv; aktuella områden; händelser och händelseförlopp; tankar, åsikter, idéer, erfarenheter och känslor; relationer och etiska frågor.

Reception

• Talat språk, även med viss social och dialektal färgning, och texter som är instruerande, berättande, sammanfattande, förklarande, diskuterande, rapporterande och argumenterande, även via film och andra medier.

Produktion och interaktion

• Muntlig och skriftlig produktion och interaktion av olika slag, även i mer formella sammanhang, där eleverna instruerar, berättar, sammanfattar, förklarar, kommenterar, värderar, motiverar sina åsikter, diskuterar och argumenterar.

• Bearbetning av egna och andras muntliga och skriftliga framställningar för att variera, tydliggöra och precisera samt för att skapa struktur och anpassa till syftet och situationen.
I detta ingår användning av ord och fraser som tydliggör orsakssammanhang och tidsaspekter.

**Kunskapskrav**

**Betyget E**
Eleven kan förstå **huvudsakligt innehåll och uppfatta tydliga detaljer** i talad engelska i varierande tempo och i tydligt formulerad skriven engelska, i olika genrer. Eleven visar sin förståelse genom att **översiktligt** redogöra för, diskutera och kommentera innehåll och detaljer samt genom att med **godtagbart** resultat agera utifrån budskap och instruktioner i innehållet.

I muntliga och skriftliga framställningar i olika genrer kan eleven formulera sig **relatirt** varierat, **relativt** tydligt och **relativt** sammanhängande. Eleven kan formulera sig **medvisst** flyt och i **någon mån anpassat** till syfte, mottagare och situation. Eleven bearbetar, och gör förbättringar av, egna framställningar.

**Betyget C**
Eleven kan förstå **huvudsakligt innehåll och uppfatta väsentliga detaljer** i talad engelska i varierande tempo och i tydligt formulerad skriven engelska, i olika genrer. Eleven visar sin förståelse genom att **välgrundat** redogöra för, diskutera och kommentera innehåll och detaljer samt genom att med **tillfredsställande** resultat agera utifrån budskap och instruktioner i innehållet.

I muntliga och skriftliga framställningar i olika genrer kan eleven formulera sig **relatirt** varierat, tydligt, sammanhängande och **relativt strukturerat**. Eleven kan även formulera sig med flyt och **viss anpassning** till syfte, mottagare och situation. Eleven bearbetar, och gör **välgrundade** förbättringar av, egna framställningar.

**Betyget A**
Eleven kan förstå **såväl helhet som detaljer** i talad engelska i varierande tempo och i tydligt formulerad skriven engelska, i olika genrer. Eleven visar sin förståelse genom att **välgrundat**
och nyanserat redogöra för, diskutera och kommentera innehåll och detaljer samt genom att med gott resultat agera utifrån budskap och instruktioner i innehållet.

I muntliga och skriftliga framställningar i olika genrer kan eleven formulera sig varierat, tydligt, sammanhängande och strukturerat. Eleven kan även formulera sig med flyt och viss anpassning till syfte, mottagare och situation. Eleven bearbetar, och gör välgrundade och nyanserade förbättringar av, egna framställningar.