Reliability of writing evaluation methods in the EFL classroom

Eric Hessels (4294831)

MA English Language and Culture: Education and Communication

Master Thesis

Utrecht University

First Assessor: Prof. dr. Huub van den Bergh

Second Assessor: Prof. dr. Aoju chen

Table of Contents

Abstract

The evaluation of writing is challenging for foreign language teachers in many different regards. Different evaluation methods have frequently been the topic of writing research, but the reliability of different evaluation methods has rarely been studied side-to-side. This study aimed to analyse the inter- and intra- rater reliability in an EFL context of three different evaluation methods: holistic, analytic, and relative evaluation. Four secondary school teachers were selected to rate twenty different written products by beginner and advanced EFL students, using every evaluation method once for every written product. Raters used an adapted version of the ESL Composition Profile for analytic evaluation, and one reference text for relative evalution. Results indicated a high degree of agreement between raters and great internal consistency for individual raters, showcasing differences between L1 and L2 writing evaluation procedures. However, no significant effects were found for differences between correlation coefficients of different evaluation methods. The reliability of individual raters and the reliability across multiple raters was not affected significantly by the evaluation method. Various explanations for these findings are discussed, together with classroom implications and recommendations for further writing studies on the reliability of evaluation methods.

Keywords: EFL, writing, reliability, evaluation, raters, holistic, analytic, relative, consistency, reference text

Reliability of writing evaluation methods in the EFL classroom

Introduction

In the context of learning foreign language writing, the importance of accurate and relevant feedback cannot be overstated enough. Accurate feedback is essential for language learners to improve their skills, learn the many nuances of a specific language, and overcome cognitive overload during writing (Bereiters & Scardamalia, 1987). At the same time, teachers frequently differ in their assessment on the same writing products. Such variability is called *rater variability.* According to L1 writing evaluation studies, there are many variables which can cause inter-rater variability, including the (educational) background of the raters, prior experience, knowledge of the raters about the writers, preconceptions about the given assignment, and of course the rating approach itself (Weigle, 2002; Malouff, Emmerton, & Schutte, 2013; Shohamy, Gordon, & Kraemer, 1992). Instructional teaching methods used at secondary schools attempt to counteract this by providing guidelines and evaluation forms for teachers, but the evaluation forms of instructional teaching methods are not always used by teachers themselves (Meestringa & Ravesloot, 2014). Rather, it is more likely the consideration of which evaluation method to use is more dependent on the aforementioned rater-related variables and practical considerations. After all, detailed personalised feedback on the writing product can be both time-consuming and challenging.

Not only is writing assessment a challenge for teachers, there are also various sources of variability which weaken the strength of individual assessments. Determining to what degree a written product showcases the writing ability of a student is challenging due to various internal factors which can influence performance. Wesdorp (1981) names several sources of variability which can make assessment more unreliable, the first of which being *participant variability*: the performance of participants will inevitably differ due to fluctuations in physical and mental capacity on a moment-to-moment basis. External factors which can distract participants during writing, like uncomfortable (class-)room temperatures can have a direct influence on their performance. These factors could be seen as more internal sources of variability. Another source of unreliability is *test variability*. This refers to the degree of variability in participant performance due to the inherent characteristics of the test itself. Different types of writing prompts for students lead to writing products of different quality simply due to their structural and contextual differences (Wesdorp, 1981; Weigle, 2002; Huang, 2009). It should be noted this is the case for L1 writing. In the context of L2 writing education, it is also important to consider that the usage of a foreign language brings about additional cognitive demands during writing for foreign language users (Tillema, Van den Bergh, Rijlaarsdam & Sanders, 2012), which might make it even more difficult to work with specific types of prompts which have unfamiliar language use in them.

Minimising some external sources of variability in writing education seems to be relatively easy: the environment and moment of assessment should be kept as stable as possible (i.e. no distractions from outside, similar timeframes for assessment, comfortable classroom temperatures, etc.). On the other hand, it can be difficult to control for the context of the prompt. The main challenge is to ensure the prompts themselves do not give specific students a large advantage due to prerequisite knowledge, and while it is preferable to have students assessed through multiple different prompts before making conclusions about a student's writing ability (Bouwer, Béguin, Sanders, Van den Bergh, 2015), it is not always feasible for teachers to give students many different assessment opportunities for writing due how time-consuming checking the writing products can be. This, however, brings back the aforementioned problem of rater variability, since the rater itself is also negatively affected by the time-related constraints. This is because in these limited moments of assessment, the rater itself comes forward as an additional and very significant source of unreliability. Raters frequently disagree among each other about the text quality, and also disagree with themselves during re-evaluation (Diederich, French, & Carlton, 1961; Weigle, 2002). It should be said such findings were mostly documented in L1 studies. Very little research has been done on the evaluation on the reliability of writing evaluation in an L2 context. Therefore, it is uncertain to what degree rater reliability in general is equally low in an L2 setup. Nevertheless, it can be assumed reliable evaluation of writing performance in the foreign language classroom is a challenging overall. Rater variability has been frequently covered in many different areas of research and has ultimately resulted in many different evaluation methods being developed. In this study, the concept of rater reliability and variability will be further defined, and an overview will be given of how the reliability of writing evaluation in the classroom is affected by the evaluation method used.

Theoretical Background

Defining reliability

When considering the variablitity or rater assessment, a distinction should be made between two types of reliability: the intra-rater reliability and the inter-rater reliability. The *intra-rater reliability* of assessment refers to the agreement between two ratings of the same rater on identical texts. If an individual rater were to be given exactly the same text on two different moments and the rater would be unable to remember the previous score, it is possible the rater's score for text quality would not be consistent. Further sources of unreliability by raters can also be seen in the assessment of multiple raters, which is also known as *inter-rater-reliability*. Teachers frequently give different scores for the same texts and also rank certain evaluation criteria differently (Wesdorp, 1981). These two elements are both referred to when defining inter-rater reliability, though they are inherently two different effects. It has been shown time and again that ratings of L1 texts vary largely over different raters, even if the evaluation criteria are kept the same (Lunz, Wright, & Linacre, 1990; Weigle, 2002). This is because teachers frequently disagree on the importance of specific sub-aspects of text quality, like structure, organization, and content. This phenomenon is called the significal effect. While it has been well-documented raters vary in how important they consider specific text traits to be for their assessment of text quality, it is generally unclear how raters arrive at certain conclusions and what exact role reading and scoring procedures play (Huot, 1990). Aside from this, intra- and inter-rater reliability are of course closely related to one another. If a rater were to be incredibly inconsistent in the assessment of very similar texts, it is likely this rater's assessment will differ from those of other raters.

Possible reliability issues during evaluation

Aside from the disagreements on how important certain text traits are for assessment, Wesdorp (1981) highlights several other effects which can be used to explain the general unreliability of teacher ratings. For one, there is the halo-effect. Prior experience with or knowledge of the student’s identity might influence a rater during assessment. For instance, Malouff, Emmerton, and Schutte (2013) have shown the performance of students during an oral presentation directly influenced the ratings of teachers who graded an unrelated written text afterwards. Other studies have also shown how biases regarding student’s identity are also the cause of unwanted variability (Nieva & Gutek, 1980; Nisbett & Wilson, 1977; Archer & McCarthy, 1988).

Another unwanted effect is the sequential effect: the ordering of texts has a notable effect on the assessment of teachers. This is one effect which directly influences intra- and inter-rater reliability, as the scores of other texts which influence a rater’s assessment at a specific moment. To illustrate this, consider the effect of a teacher rating five texts in succession. If the first four texts were disappointing in terms of text quality, it is very likely an ‘average’ text will be rated significantly higher than when it is assessed in isolation from other texts. The inverse can be the case as well: if four outstanding texts were to be graded first, an average text is very likely to be graded lower than when graded in isolation. Sequential effects have frequently been reported in for the assessment of writing and are prevalent in different levels of education (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes, Keeling & Tuck, 1980; Speaer, 1997; Attali, 2011).

Finally, shifting norms between raters should also be considered. Depending the attitude of the rater, the distribution of high grades and low grades can differ. For instance, while one rater might argue a maximum score is only attainable under the rarest of circumstances for excelling students, others might give a maximum score more readily. There are several different solutions to each source of unreliability. The sequential effect, for instance, might be reduced by using reference texts for comparison. To combat the signical effect and shifting norms, the reliability of assessment can be positively influenced by clear assessment criteria and rater training (Lumley, 2002). Of course, to have clear assessment criteria means these criteria are to be defined based on the original prompt. A summarised overview of all the aforementioned effects, together with several different solutions, can be found in Appendix A.

Assessment Methods

Overall, it is clear there are many possible effects which can occur during the evaluation process which ultimately influence reliability of raters. In general, it seems carefully controlled rating procedures and a well-defined assessment method greatly facilitate the rating procedure. The question is which assessment method should be used. There are, after all, a plethora of different evaluation methods, which can ultimately influence the reliability of assessment to a significant degree (Van den Bergh, De Maeyer, Van Wijen, & Tillema, 2012; Schoonen, 2005). There are many different methods of assessment which could be named. For now, four types of evalution will be considered due to their frequency of use and possible applicability for classroom practice: holistic evaluation, analytic evalution, primary trait evaluation, and relative evalution.

Holistic evaluation is perhaps the most common means of evaluation. It involves using the general judgement of the teacher of the whole text to determine the level of the students' writing. This method of evaluation is time-efficient, and its usefulness could be substantiated by arguing the inherent qualities of a written text cannot be assessed and quantified through objective criteria (Hamp-Lyons, 1990). On the other hand, an assessment method such as holistic evaluation would provide raters the required freedom to assess a text properly. Looking at foreign language writing assessment, evaluating the text as a whole would also allow raters to be more flexible in the considerations they make when assessing the mistakes of non-native speakers rather than native speakers mistakes (Oller and Perkins, 1980; Jacobs et al, 1981; Hamp-Lyons, 1990). Yet, this method of evaluation is heavily dependent on a teachers' experience and approach, and multiple studies have showcased the general unreliability of teachers’ holistic assessment of student texts (Diederich, French, & Carlton, 1961;Weigle, 2002). Wesdorp (1981) also notes holistic assessment with a single rater has a lower inter-rater reliability than holistic assessment by multiple raters, suggesting jury-assessment might resolve the reliability-related shortcomings. At the same time, Pilliner (1969) also criticises the inherent validity of such a solution, since stable raters of a jury who heavily disagree on a certain text would simply cancel each other out. The true qualities of a text then would not be reflected properly by its assessment (Wesdorp, 1981).

Another method which can be used is *primary trait evaluation*. This method of evaluation involves awarding a score to a writing product based on to what degree it satisfies the characteristics of a certain prototypical text. For instance, for an academic essay, a rater might choose argumentative reasoning as the primary trait. This would mean the rater would only evaluate whether the writing product has logically sound argumentative constructions which aim to convince readers of a certain point, leaving aside other textual or structural elements. While this allows for specific learning goals and simplifies evaluation overall, it comes with the disadvantage that relevant aspects of a writing product might not be considered during the grading process (Schoonen & De Glopper, 1992). To come back on the previous example, it would be considered unusual for an academic essay to be given a perfect grade because of its solid argumentation, even though the essay lacks appropriate academic register or consistently uses specific tenses erroneously. Of course, it is ultimately the severity of such errors which affect the degree the quality of the writing. Llach (2007) noted the most minor type of lexical errors, misspellings, impacted overall writing quality the least. On the other hand, certain measures of lexical richness, like lexical diversity and lexical originality considerably influenced the quality of written products.

Analytic evaluation is another way to evaluate writing skill. While the term is also frequently used in research do describe the process of assigning separate scores to different text criteria (e.g. grammar, structure, textual characteristics, etc.), it can also be used to describe the process of evaluators using a prompt-specific ruleset to deduct or reward points based on several different criteria. For instance, an analytic evaluation model might describe that conjugating a tense incorrectly in a specific written product costs three points per mistake, up to a maximum of a 20-point deduction for the total score to determine the final score. With analytic evaluation, it is therefore possible to be incredibly specific with the description of evaluation criteria, and therefore also provides more useful information for students as well. This evaluation method is generally considered to be more precise and reliable because of more specific guidelines. Several studies have shown the evaluation method to have a higher interrater reliability in comparison to holistic and primary trait evaluation (Hamp-Lyons, 1991; Jacobs et al., 1981; Bachman & Palmer, 1996). However, the guidelines on their own do not simply increase the reliability of assessment. Carlson et al. (1985) reported high correlations between different raters who were trained and standardised on a daily basis, and similar procedures are often done with analytic evaluation schemes to ensure agreement. In a series of studies, Meuffels (1994) illustrated analytic evaluation was not necessarily more reliable than holistic evaluation. In addition, guidelines can also become restricting in an attempt to produce reliable evaluation, which undermines the validity of the evaluation method. Lumley (2002) and Wesdorp (1981) showed the rules and guidelines are unable to cover all eventualities and usually do not take the relationships between individuals' traits into account. In addition, Lumley states even when using an analytic evaluation model which attempts to describe rating procedures as accurately as possible, raters are nevertheless influenced by the complex intuitive impression of the text upon first reading it (Weigle, 2002). The ways in which raters resolve these two challenges with analytical evaluation models are ultimately quite indeterminate, and rigorous training using the analytic model is required to overcome this pitfall. This is because analytic evaluation might cause raters to focus on lower order mistakes rather than higher-order mistakes, which detracts from the construct validity and generalisability of the evaluation method (Tillema, Van den Bergh, & Rijlaarsdam, 2012; Tillema, 2012).

Finally, it is also possible to directly compare different writing products. *relative evaluation* involves comparing a writing product to other texts to determine the quality of the text. Blok (1985), Purves (1992) and Schoonen (2005) have described this as a more reliable form of evaluation in comparison to holistic and analytic evaluation because it avoids sequential effects and shifting norms. Using a single continuous frame of reference, teachers are therefore able to analyse a text without being influenced by the preceding texts. One way to implement this type of evaluation is to construct a rating scale with several different texts by asking a team of experts to pick out a prototypically average text from a large sample, and then to ask raters to directly compare other texts to the prototypically average text (See Van den Bergh, De Maeyer, Van Weijen & Tillema, 2012 for an example). By determining which texts are consistently rated higher (e.g. consistently rated twice as good) in comparison to the preceding text, a scale with several different 'anchor texts' can be constructed. However, it should be noted this method of evaluation is time-intensive and is impractical for most classroom writing assignments due to the construction of the scales, as it involves a heavily controlled methodology. That said, the rating scales can be used for assignments of the same genre. For standardised assignments, this form of evaluation is considered the most reliable method of evaluation (Tillema et al., 2012). A simpler variant of this type of evaluation is to pick out a prototypically average text, and to let raters only use this text in their evaluation. Such an approach can also produce reliable ratings (Kox & Van Den Bergh, 2018).

Overall, it is clear that relative evaluation is well-supported in literature as a reliable means of evaluation of texts in comparison to holistic evaluation. Wesdorp (1981) stated too much freedom for raters during the evaluation can be detrimental to the inter-rater reliability, though specific and strict guidelines make them more challenging to use. The consistency and inter-rater reliability are higher for analytic assessment than for holistic assessment. That said, training raters properly and using a straightforward analytic assessment model might further increase reliability. Earlier studies have mainly compared the writing process and quality of writing products between L1 and L2 (Tillema, 2012), or have compared only analytic and holistic evaluation (Weigle, 2002). Pollmann, Prenger, and De Glopper (2012) concluded in their study using relative evaluation produced high interrater agreement and jury reliability scores but did not compare it to other forms of evaluation. In addition, a total of 100 different texts and seven different experts were used to select the anchor texts for the rating scale, which far supersedes any practical means of evaluation in teaching practice. Rather, using a single reference text might be more practical for assessment, yet also might give some of the benefits commonly associated with relative evaluation. Comparing different methods of evaluation for L2 writing might give insight into the overall usability of those different methods of evaluation. Based on this, the following research question can be formulated:

To what degree does the method of evaluation influence the overall reliability of rater scores in L2 writing education?

Based on earlier studies, it is unsure which evaluation method will be the most reliable. Previous studies have both supported and dismissed an increase of reliability when using analytic evaluation methods. Using relative evaluation might increase the reliability of assessment in comparison to holistic evalution, though direct side-by-side comparisons in an L2 context have not been done before.

Method

Participants

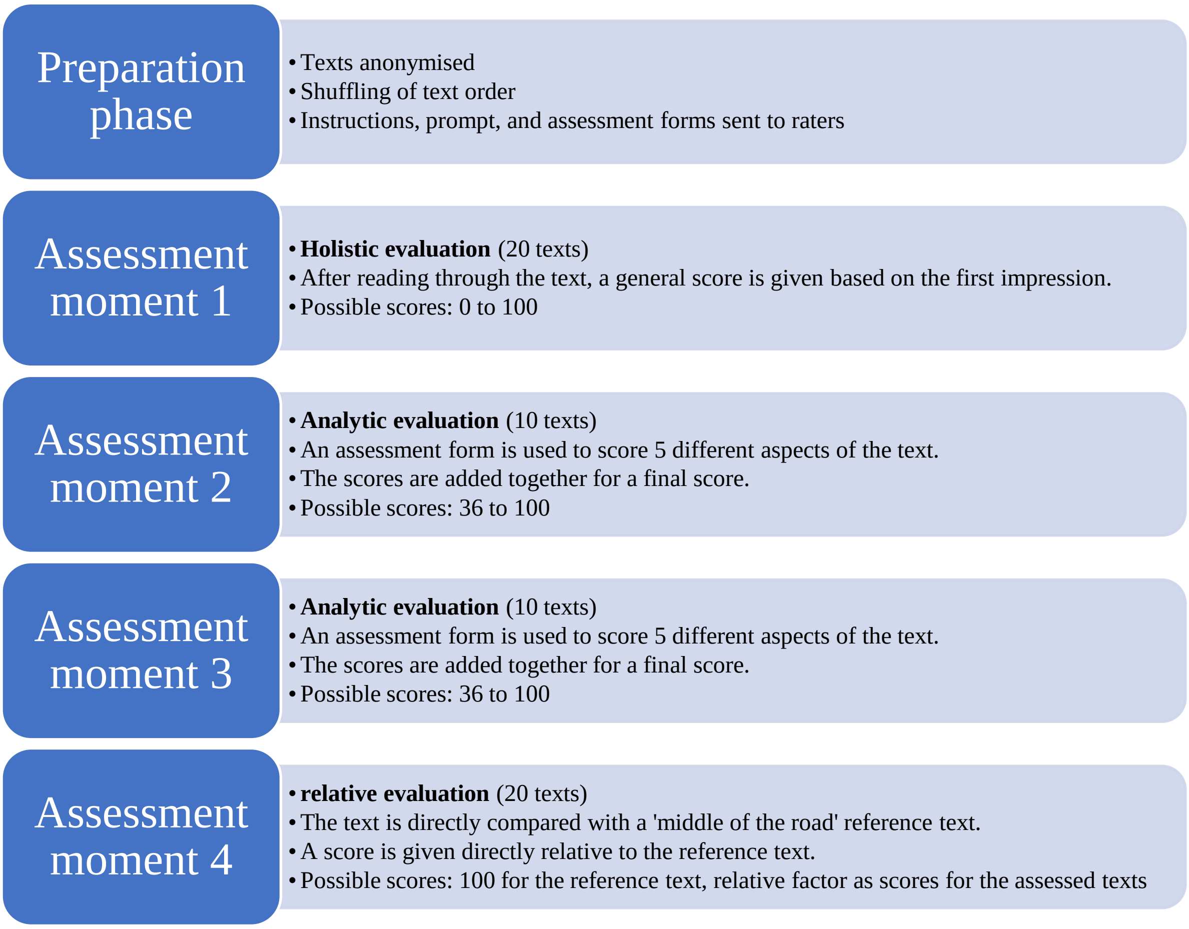
For this study, four teachers from two Dutch secondary schools, including the researcher of this study, were selected to assess the writing quality of 20 different student texts. Regarding experience, teacher A has been teaching for 2 years in secondary schools, teacher B for 4 years, teacher C for 5 years, and teacher D for 3 years. Due to practical restraints, it was not possible to recruit more experienced teachers. For the written products, 10 students from second-year havo/vwo classes and 11 students from fourth- and fifth-year vwo classes were asked to write a short persuasive text. The difference in level was set up to function as a ‘known-groups method’ to allow for some underlying reasoning about the construct validity of the assessment afterwards. Only the researcher of this study had had experience teaching some of the students before. Because of this, the writing texts were made anonymous. An external participant replaced any contact information used by the students in the written products with fake addresses and replaced included student names with the name ‘Anna Johnson’. One written text was excluded after anonymisation by the researcher of this study to use as a reference text for one of the evaluation methods (see methods of evaluation for more information).

Procedure

Students were asked to write a short persuasive letter during one 60-minute lesson based on the well-renowned ‘Smikkel Case’, designed by Rijlaarsdam and Braaksma (2004). It was decided to use a prompt which would require students to write a short persuasive letter because the general requirements could be simply formulated for students. The assignment involved writing a short response letter with a specific request to participate in a special promotional contest of a chocolate bar producer to win two tickets for the London Theatre. Students had to write a persuasive formal letter which would convince the company to let them participate in the contest with only eight out of ten of the ‘required tokens’ for the promotional contest, and two ordinary wrappers. The main assignment given to students was to ensure the chocolate company would nevertheless send the tickets. The address information of the recipient was also given in the prompt. Students were explicitly asked to write the letter in English. The full prompt can be found in appendix B. The texts were written during classroom hours to ensure equal testing conditions for the individual participants. During this time, students used a computer with a word processor (with the language checker turned off) to write the response letter. Afterwards, students were given a small present to reward them for their participation.

Afterwards, every rater was asked to check the 20 written products, using the same three evaluation methods for every text: holistic evaluation, analytic evaluation, and relative evaluation. Primary trait evaluation was left out because this method of evaluation inherently aims to score a more specific trait of writing rather than overall text quality, which would make comparisons between scores problematic. The raters were instructed to score the texts using every method of evaluation once, and used one evaluation method at a time. For analytic evaluation, raters were advised to space out assessment into two separate moments. his was done because of the time-consuming nature of analytic evaluation and to prevent mental fatigue from hindering assessment.

The raters used the evaluation methods in different orders and went through all texts with one specific evaluation method at a time, using the same ordering of texts. However, raters always used holistic evaluation before relative evaluation, since the sample text might influence the judgement of the raters. For instance, one rater was instructed to first use holistic evaluation to score all the texts, then to do the same with analytic evaluation, and finally to use relative evaluation. The varying orders between raters prevented a specific order used for assessment from influencing all the raters and allowed for direct comparisons between the different assessment orders. In figure 1, an overview of the rating process can be seen.



Preparation phase

- Texts anonymised
- Shuffling of text order
- Instructions, prompt, and assessment forms sent to raters

Assessment moment 1

- **Holistic evaluation** (20 texts)
- After reading through the text, a general score is given based on the first impression.
- Possible scores: 0 to 100

Assessment moment 2

- **Analytic evaluation** (10 texts)
- An assessment form is used to score 5 different aspects of the text.
- The scores are added together for a final score.
- Possible scores: 36 to 100

Assessment moment 3

- **Analytic evaluation** (10 texts)
- An assessment form is used to score 5 different aspects of the text.
- The scores are added together for a final score.
- Possible scores: 36 to 100

Assessment moment 4

- **relative evaluation** (20 texts)
- The text is directly compared with a 'middle of the road' reference text.
- A score is given directly relative to the reference text.
- Possible scores: 100 for the reference text, relative factor as scores for the assessed texts

Figure 1. Summary of the step-by-step process of the rating procedure. The order ‘holistic-analytic-relative’ was used in this example, though other orderings were used as well.

Methods of Evaluation

The four raters were asked to use every method of evaluation once for every text for equal comparisons between raters. The four raters were also asked to space the moments of evaluation between different methods. For *holistic evaluation*, the raters were asked to read through the entire text once and give a grade on a 1-to-100 scale for all three categories.

For *analytic evaluation*, the raters were asked to grade the texts based on an assessment form which describes several important traits for the assignment. The assessment form used for this study was adapted from ESL Composition Profile (Jacobs et al., 1981), which can be found in table 1. While originally used for the assessment of academic writing, it was decided to use this assessment form for three reasons. For one, the assessment categories were concise, clear, and were well-differentiated from one another. second, the assessment form was easily adapted to suit the persuasive writing prompt used in this study. Finally, it saved time for the participants to use this evaluation form in comparison to more detailed evaluation forms which ask raters to count specific mistakes. The assessment profile is divided into 5 different categories: *content, organisation, language use, vocabulary, and mechanics*. Every category can be scored based on four different subcategories: *excellent to very good, good to average, fair to poor,* or *very poor*. Every subcategory has its own score range. For instance, a rater can award 26 to 22 points if it believes the content of a text to be good to average. It should be noted the scores for every category of this rating profile does not have 0 as a minimum score: this results in the minimum score being at least 36. Therefore, the possible range of the scores for this evaluation method is different from holistic scores. That said, this did not affect the assessment of reliability per rater, since every method was used by every rater, allowing for direct comparisons between raters. Unfortunately, due to time constraints and the distance between the two schools, it was not possible to assign time to train the raters with the ESL composition profile.

For *relative evaluation,* one ‘benchmark’ text was selected from the sample which functioned as a model of reference. In figure 2, the benchmark can be found. The benchmark text was selected by the researcher of this study with the help of one other English teacher. Both the researcher of this study and the reference English teacher selected one ‘average’ text from the writing text samples. The text was selected such that it was as close as possible to being average in every possible respect in comparison to the sample, with the underlying reasoning being summarised. The same text was ultimately selected by both the researcher of this study and the reference English teacher.

Anna Johnson
Fake address 123
1234 AB The Hague
The Netherlands

Yum-Yum inc. Marketing Department
New Headway Street 33
1023AB, London
United Kingdom

Date: 4 may 2019
Subject: Tickets for the London Theatre

Dear sir/madam,

My name is Anna. I live in the Netherlands. I am writing this letter because I have a request for the contest about your Yum-Yum chocolate bars. There is a problem when I wanted to enter the contest. I have 8 entry tokens but I cannot find anymore of them in store. I still want to participate.

I really want to participate because I really like London Theatre and I also like Yum-Yum chocolate bars. I think they taste great. In the envelope you will find 2 normal wrappers and 8 entry tokens. So I tried to get all the tokens I need but there were no more in stores in The Netherlands. The promotional offer runs until June 26th, 2019, and right now it is may so I hope I can still participate.

I really love Yum-Yum chocolate and really want to see the play. Please let me enter. Thank you for the your assistance.

Kind regards,

Anna Johnson

Figure 2. Reference text, written by a vwo 4 student.

	LEVEL	**Explanation of each level**
SCORE	**MAX - MIN**	
CONTENT	**30 - 27**	**EXCELLENT TO VERY GOOD:** Knowledgeable; substantive; thorough development of argument; relevant to assigned topic
	26 - 22	**GOOD TO AVERAGE:** Some knowledge of subject; adequate range; limited development of argumentation; mostly relevant to topic, but lacks detail
	21 - 17	**FAIR TO POOR:** limited knowledge of subject; little substance; inadequate development of argument
	16 - 13	**VERY POOR:** Does not show knowledge of subject. non-substantive. not pertinent. OR not enough to evaluate
ORGANISATION	**20 - 18**	**EXCELLENT TO VERY GOOD:** fluent expression; ideas clearly stated/supported; succinct; well-organized; logical sequencing; cohesive
	17 - 14	**GOOD TO AVERAGE:** somewhat choppy; loosely organized but main ideas stand out; limited support; logical but incomplete sequencing
	13 - 10	**FAIR TO POOR**: non-fluent; ideas confused or disconnected; lacks logical sequencing and development
	9 - 7	**VERY POOR:** does not communicate; no organization; OR not enough to evaluate
VOCABULARY	**20 - 18**	**EXCELLENT TO VERY GOOD:** sophisticated range; effective word/idiom choice and usage; word form mastery; appropriate register
	17 - 14	**GOOD TO AVERAGE:** adequate range; occasional errors of word/idiom form, choice, usage *but meaning not obscured*
	13 - 10	**FAIR TO POOR:** limited range; frequent errors of word/idiom form, choice, usage; *meaning confused or obscured*
	9 - 7	**VERY POOR:** essentially translation; little knowledge of English vocabulary, word form; OR not enough to evaluate
LANGUAGE USE	**25 - 22**	**EXCELLENT TO VERY GOOD:** effective complex constructions; few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
	21 - 18	**GOOD TO AVERAGE:** effective but simple constructions; minor problems in complex constructions; several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*
	17 - 11	**FAIR TO POOR:** major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; *meaning confused or obscured*
	9 - 7	**VERY POOR**: major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; *meaning confused or obscured*
MECHANICS	**5**	**EXCELLENT TO VERY GOOD:** demonstrates mastery of conventions; few errors of spelling, punctuation, capitalisation, paragraphing
	4	**GOOD TO AVERAGE:** occasional errors of spelling, punctuation, capitalisation, paragraphing *but meaning not obscured*
	3	**FAIR TO POOR:** frequent errors of spelling, punctuation, capitalisation, paragraphing, *meaning confused or obscured*
	2	**VERY POOR:** no mastery of conventions; dominated by errors of spelling, capitalisation, paragraphing; OR not enough to evaluate
TOTAL SCORE		

Table 1. Adapted ESL Composition Profile.

The underlying reasoning was then summarised in the form of a brief description. This description, together with the reference text, was given to teachers for the relative evaluation procedure so the raters had contextual information about the judgements made during the selection phase. Raters were then asked to assess the overall quality and structure of the texts by directly comparing them to the reference text. A score of 100 was given to the benchmark essay, and raters were asked to score the other texts based on this score. For instance, a score of 200 would imply the text would be twice as good as the reference text.

Analysis

For the analysis, the scores of text quality were the key variable. In some analyses, the focus was on differences in mean scores between raters, rating methods and grade of writers. In other analysis the focus was on Pearson correlations coefficients between raters, rating methods and scores of writers. for the scores of different teachers for the same assessment method were calculated to measure inter-rater reliability, while internal comparisons of correlations between the scores of individual teachers were used to measure intra-rater reliability. The Spearman-Brown prophecy formula (Salkind, 2010) was used for homogeneous test extension to estimate jury reliability.

Results

In table 1, the mean scores per evaluation method are presented. The standard deviation illustrates the large variation in scores given per text. For holistic evaluation, mean scores were between 40.3 (SD = 26.9) and 61.7 (SD = 15.3). The scores for raters had a possible range of 1 to 100, though no scores higher than 90 were given. Interestingly, teacher C gave no score with holistic evaluation lower than 40. For analytic evaluation, the mean scores were between 60.7 (SD = 18.6) and 66.8 (SD = 20.1). For this evaluation method, the possible range was 36 to 100 because of the difference in scoring method for the ESL Composition Profile. In practice, the scores of raters for this evaluation method were between 36 and 94. Finally, the scores relative evaluation showed the largest degree of variance. The mean scores for this evaluation method were close together and between 68.4 (SD = 51.1) and 71.8 (SD = 57.0). For this evaluation method, there was no possible maximum or minimum given due to the nature of the evaluation method. It is immediately noticeable based on the standard deviations that there was large variance between different raters, regardless of the evalution method.

Table 1.

Average scores per evaluation method (standard deviation between parentheses).

	Teacher A		Teacher B		Teacher C		Teacher D	
	M	SD	M	SD	M	SD	M	SD
Holistic	47.8	26.3	40.3	26.9	61.7	15.3	48.8	28.3
Analytic	60.7	18.6	62.1	17.1	65.1	18.3	66.8	20.1
Relative	71.8	57.0	68.8	52.2	75.7	50.1	68.4	51.1

Paired samples t-tests were done per rater to determine significance for the differences between the mean scores per evaluation method. An alpha level of .05 was used for all tests. If holistic and analytic ratings are to be compared, a significant difference in mean scores can be found for three out of four teachers (t (19) ≥ 5.9; p < .001). If holistic and relative scores are to be compared, a significant difference in mean scores can be found for the same three teachers (t (19) ≥ 3.2; p ≤ .004), whereas for comparisons between analytic and relative did not result in any significant differences. In appendix C, a full overview of all the test statistics can be found.

Minimum and maximum scores for holistic (1 to 90), analytic (36 to 95), and relative (0 to 170) also indicate raters found some texts to be of very high quality and others of very poor quality. This is to be expected, since some of the variance between the mean scores can also be accounted for by differentiating between the scores of the two student categories. The group of writers consisted of 'beginner' EFL students and advanced EFL students. In table 2, the mean scores per student group are presented. A significant difference in level was found when taking the average scores of raters between the two groups of students for holistic evaluation , t (18) = 10.7, p <.001, analytic evaluation, t (18) = 11.4, p <.001, and relative evaluation, t (18) = 11.8, p <.001. With the average taken from all raters, scores between the beginner and advanced students differed 42.9 points for holistic evaluation, 33.8 points for analytic evaluation, and 94.1 points for relative evaluation.

It should also be noted that due to the different scoring ranges and methods, a direct one-to-one comparison is not easily possible between the scores of different evaluation methods. For instance, a score of 55 with holistic evaluation might be considered a 'middle of the road' score. A score of 55 with holistic evaluation would not be comparable to a score of 55 for analytic evaluation. Instead, a score of 71 would instead be a more similar score, if the same reference point of 55% is chosen across the possible range.

Table 2.

Average scores per evaluation method per student year (standard deviation between parentheses).

	Teacher A		Teacher B		Teacher C		Teacher D	
Year	2nd	4/5th	2nd	4/5th	2nd	4/5th	2nd	4/5th
Holistic	23.5 (10.0)	72 (10.3)	16.6 (12.9)	63.9 (12.5)	48.3 (9.6)	75 (4.7)	24.2 (16.4)	71.1 (11.0)
Analytic	43.4 (6.1)	78 (7.5)	47.4 (10)	76.9 (6.8)	47.9 (6.7)	82.2 (6.4)	48.4 (12)	85.1 (7.3)
Relative	18.5 (17.8)	125 (22.5)	20.5 (15.6)	117 (23.3)	34.5 (13.1)	116.5 (38.5)	22.7 (16.2)	114 (28.2)

Of course, for relative evaluation, the reference score in this specific comparison would be 100, since the reference text was originally given a score of 100 before the rating phase.

In table 3, the mean correlations and corresponding jury reliability can be found. When looking at the correlations between raters for the same evaluation method, the mean correlation coefficient across all raters was incredibly high for holistic evaluation ($r = .91$), as were the mean correlation coefficients for analytic evaluation ($r = .94$) and relative evaluation (.89). Consequently, jury reliability also was incredibly high for holistic ($\rho = .97$), analytic ($\rho = .98$), and relative evaluation ($\rho = .97$). This indicates there was a high degree of agreement between raters when looking exclusively at the same evaluation method.

Table 3.

Mean correlations and jury-reliability per evaluation method.

Method	r (mean)	jury-reliability
Holistic	.91	.97
Analytic	.94	.98
Relative	.89	.97

To compare the different coefficients and to determine whether the differences between the correlations for the juries were significant, a Feldt test for comparisons between small samples of raters was used (Feldt & Kim, 2006). No significant difference was found for the mean correlation across all raters for holistic evaluation and analytic evaluation ($F_{15,15} = 0.680$, $p > .05$). This was also the case for the mean correlations for holistic and relative evaluation ($F_{15,15} = 1.32$, $p > .05$) and the mean correlations for analytic and relative evaluation ($F_{15,15} = 1.96$, $p > .05$). This indicates there seemed to be no differences between the jury reliabilities of different evaluation methods.

Of course, coefficients can also be calculated across different evaluation methods. In the case of a comparison between holistic and analytic scores, the scores for holistic can be compared to the mean scores for analytic evaluation to calculate a correlation coefficient for every individual rater. Taking the average from these correlations results in a correlation between juries. In table 4, the correlation coefficients between the three ratings methods are shown. The correlation coefficients for these scores were corrected for attenuation to account for possible inaccuracies of measurement (Spearman, 1904).

Table 4.

Correlations between three rating methods as observed (below the diagonal) and corrected for unreliability (above the diagonal).

	Holistic	Analytic	Relative
Holistic	-	1.00	1.00
Analytic	0.98	-	.99
Relative	.97	.97	-

Very high correlations er found between holistic and analytic ($r = 1.00$), between holistic and relative ($r = 1.00$) and between analytic and relative ($r = .99$). However, using the same test as before, no significant difference between the holistic-analytic and holistic-relative correlations was found ($F_{15,15} = 1.89$, $p > .05$). No significant difference between the holistic-analytic and analytic-relative coefficients were found as well ($F_{15,15} = 2.08$, $p > .05$). Finally, no significant difference between the holistic-relative and the analytic-relative coefficients were found either ($F_{15,15}$ = 1.1, p > .05). This means when taking the four raters as a jury, the order of the texts does not differ as a result of the evaluation method. In other words, the order for the texts used by raters for holistic evaluation was functionally identical to analytic and relative evaluation.

Discussion

This exploratory study aimed to assess differences in reliability between three evaluation methods: holistic evaluation, analytic evaluation, and relative evaluation. Previous studies have showcased the frequent problem of high rater variability among different evaluation methods. the results of this study instead showed high inter- and intra-rater reliability for all evaluation methods, though the differences in agreement between different evaluation methods failed to reach significance. That said, several differences in method should be pointed out here. For one, previous studies have mainly used L1 writers of a higher level. In such a context, raters have to consider more subtle matters during the evaluation of texts. In addition, while many L1 studies have used *homogenous* groups of L1 writers, this study used a *heterogenous* group consisting beginner and advanced L2 writers. Correlations between groups of the same level would be lower than the overall correlations, because the overall correlations take the differences between student levels into account. These differences in setup make comparisons between earlier studies more difficult.

Overall, based on the scores given by raters, it is clear the raters were frequently in agreement using the same evaluation method. Comparisons between different evaluation methods showed large differences in mean scores. This could be partly attributed to the differences in scale for the different evaluation methods, which also affects how the differences between raters and groups were expressed. These groups were also significantly different from one another based on the scores of raters, clearly showcasing the difference in writing ability. Regarding reliability, every evaluation method had high correlations between jury members, suggesting high jury-reliability overall. When comparing the jury scores of different evaluation methods, high correlations were found again, suggesting the evaluation methods measured writing ability similarly, though with a different rating scale. It is therefore possible the evaluation method might not be as impactful for L2 writing evaluation as is the case with L1 writing evaluation. This will be further discussed below, together with other findings.

As mentioned before, there was a high degree of agreement between raters, regardless of the evaluation method, though the differences failed to reach significance. Disregarding the lack of significant differences between correlations, it was unusual to have such high correlations between raters. After all, the overall reliability scores in larger-scale studies were reported to be quite low for holistic evaluation (Wesdorp, 1981; Diederich, French, & Carlton, 1961). There are several possible explanations for this. For one, it is very much possible the sample of essays were easily distinguished from one another in terms of quality. Another explanation could be that the individual raters were simply like-minded in the assessment of certain textual qualities, causing high agreement how a score would be affected if a specific type of error would be made by students. This would prevent any significal effects as described by Wesdorp (1981). Of course, the large score range and differences between individual scores per raters seemed to be mostly caused by a significant level difference.

Ultimately, the differences in reliability per evaluation method were not significant overall, though the correlations were incredibly high, suggesting there are most definitely differences between L2 writing evaluation and L1 writing evaluation. Some general observations can be made about the differences in scores. First, it was quite unusual for the high degree of agreement between raters, since the raters were not trained to use analytic and relative evaluation, and were, on average, inexperienced FL teachers. Linn and Burton (1994) originally found high agreement between raters who were extensively trained with a well-defined rubric. interrater reliability was negatively affected by ambiguity or general unclarity in scoring rubrics. Looking back at the rating rubric used by the teachers, there seemed to be very little ambiguity between teachers. However, when looking at the individual scores, it can be seen the scores given by teachers for some texts differed significantly. Such differences were even more clearly pronounced with the holistic scores. It is possible the significal effect was at hand here, which states that if multiple raters are asked to judge the same writing product without any further instruction, it is likely they will use varying approaches during evaluation due to different personal standards. Of course, this is most commonly seen with holistic evaluation due to the personal freedom in the rating approach. Nevertheless, the overall correlations were evidently close to the possible maximum, especially so when accounted for possible errors in measurement. Part of this inconsistency can be explained by the observation that there were some teachers who gave no scores lower than 40 with holistic evaluation. This was most likely because analytic evaluation was done by this rater before holistic evaluation, resulting in the teachers possibly misunderstanding the minimum and maximum possible scores for this evaluation method, creating an elevated minimum score for for holistic evaluation.

Aside from the misunderstanding regarding the possible minimum score for holistic scores, it was also noted upon inquiry by certain raters that the relative evaluation method was surprisingly difficult due to the general unfamiliarity with the rating process and the lack of multiple reference points. Indeed, if a rating scale was used rather than a reference text when placing high- or low-quality texts at the extreme ends (see Van den Bergh, De Maeyer, Van Weijen, & Tillema 2012), the rating process could have been easier for the raters. While it might be easy to grade a text which has a slightly lower or higher text quality, it can be difficult to place different texts on the lower or higher end without cross-referencing previous evaluations. Some text were considered to be clearly superior in terms of overall text quality in comparison to the reference text, but raters disagreed on the *extent* to which the texts surpassed the reference text. Large differences in scoring of texts of advanced students using relative evaluation show that not all detrimental effects are completely counteracted by using a reference text.

Regarding the order of evaluation methods, the high degree of correlation between the four raters suggests the order of rating for the evaluation methods mattered very little for the agreement between raters. Of course, the gaps in-between the assessment moments might have allowed raters to disregard their earlier ratings, only remembering their earlier approximate impression of the texts. In addition, raters were explicitly asked to space out their moments of assessment, used a separate scoring form at a different assessment moment to input the scores, and were asked to ignore their earlier scoring with other methods of evaluation. It is difficult to determine based solely on the quantitative data to what degree the earlier scores affected the judgement of raters, since no retests using the same method were done in this study. Lumley (2002) used think-aloud protocols to let raters describe their rating process during the assessment of a set of scripts. Other ways to achieve this would be asking raters to summarise their reasoning per rating category. Such procedures might prove especially useful to determine how the analytic scoring rubric is interpreted and used. Furthermore, in classroom situations, the explanations regarding the rating process would also provide an additional layer of formative feedback, which can help during the learning process (Weigle, 2002). On the other hand, this would add an additional element of qualitative analysis, further increasing the time required for rating.

Aside from the aforementioned difference in level between students in the current study, the EFL context might have also increased the overall cognitive load for beginner students in comparison to the advanced students due to their inexperience with the language. Rating the written products might have been more difficult for raters if the texts belonged to students of several different skill levels. That said, while the participants from the advanced classes seemed to be familiar with formal writing procedures based on the text characteristics and overall structure, there was a clear difference in argumentation, sentence structure, and writing style. To illustrate this, consider the following two fragments of two students:

> From: Anna Johnson
> Address: Fake Adress 123
> Postal code: 1234 AB
> City: The Hague.
>
> Dear Yum-Yum Marketing Department,
>
> As a fellow Yum-Yum lover I hereby ask you to lend me your ear. As you may know, a collection period that your department has started will end very soon. I as a true fan and lover have been collecting wrappers with tokens in it while enjoying the sweet chocolate through the journey of collection. …

> Anna Johnson
> Fake address 123
> 1234 AB The Hague
> The Netherlands
>
> Yum-Yum inc. Marketing Department
> New Headway Street 33
> 1023AB, London
> United Kingdom
>
> Date: 4/3/2019
> Subject: Tickets for the London Theatre
>
> Dear Mr./Mrs.,
>
> My name is Anna and I live in The Netherlands. I write this letter to explain my situation. As you know your company has organised a campagne. Basically you need to save up tokens that you can win by buying a chocolate bar. …

As can be seen, the overall structure of the opening of these texts of is quite similar. The sender's information is mentioned, and the first part of the letter opens with a formal greeting. Both letters also use the first paragraph immediately introduce the reason for writing (i.e. the promotional contest), though the overall writing style and sentence structure is quite different. Student one used idiomatic expressions and stylistic vocabulary to explain their reason of writing, while student two instead used simple and concise sentences. Ultimately, all written products by advanced students included address information and were structured into well-differentiated paragraphs, while no beginner students inserted the same information. Nine out of ten of the advanced students used a formal greeting and sign off, while only one beginner student did so as well. This illustrates the advanced students, in contrast to the beginner students, were aware of the general writing conventions associated with the prompt, but that the degree of proficiency regarding sentence structure, vocabulary, and argumentation were the distinguishing factors for the advanced students. The difference in ability is seen clearly as well when looking at the mean scores for the two different groups.

Even though both inter- and intra-rater reliability seemed to be incredibly consistent, it is difficult to make any conclusions about the validity of assessment. What can be said with certainty is that the difference in level was recognised by raters and caused texts of different student levels to be assessed accordingly. Ultimately, the analysis of reliability in writing education exclusively deals with the consistency and stability of the assessment by the raters. The *validity* of assessment (i.e. to what degree the testing accurately measures the skill it is supposed to measure) is left out of consideration. These two aspects of test evaluation have a paradoxal relationship which is summed up well by McColly (1970): “If a test is totally non-valid, but is nonetheless reliable in all respects, it may be worse than useless because of the likelihood of its misuse” (p. 149). Regardless of the evaluation method, it is uncertain to what degree the assignment measured the writing skill of students based on the above results. Instead, it is certain the scores given by raters in this study were quite similar to one another, and the scores given by individual raters were similarly high or similarly low, regardless of the evaluation method. Wesdorp (1981) addressed the inverse relationship between reliability and validity by highlighting how incredibly reliable methods of evaluation, combined with incredibly strict writing prompts, can limit the degree of freedom for assessment, which is to the detriment of the internal validity of the test. On the other hand, having a high degree of freedom for both the rater and the student creates a situation in which reliability of assessment becomes unreliable due to the many possible ways the writing process might go. This situation is equally unwanted, since unreliability is to the detriment of internal validity: a writing test can never be a valid way of measuring writing quality if the scoring is inconsistent.

Implications for practice

Based on the researched literature and the results from this study, several recommendations can be given to teachers who are teaching writing to ensure optimal reliability for assessment. For one, regardless of the evaluation method used, clear, descriptive evaluation criteria and prior agreement among raters on specific guidelines are essential to ensure consistency between different raters. Language teachers should discuss the evaluation format used beforehand and bring forward difficult cases during the evaluation process. This is especially true for foreign language teachers, since students might make more lower-order mistakes (i.e. spelling and grammar errors) due to their general inexperience with the language. Second, if using analytic evaluation, the rubric should use only a few rating categories with suggestions for scoring per category. Wesdorp (1981) recommends using approximately five different rating categories with concise, yet well-defined rating criteria to streamline the overall rating process. Training raters to use the rating rubric might make assessment more consistent, though raters should be cautious of being trained to consider only superficial or lower-order mistakes during assessment, as this has negative effects on the internal validity (Tillema, Van den Bergh, & Rijlaarsdam, 2012; Tillema, 2012). Third, sample texts on a rating scale can be used to avoid negative sequential effects during the rating process. Negative sequential effects can be avoided by continuously comparing the texts which are to be assessed to sample texts with a clearly defined score.

Conclusion

To sum up, this study aimed to compare three different methods of evaluation in foreign language writing education. Based on the results of a small-sample exploratory study with four secondary school teachers and twenty written texts, no significant differences in inter- and intra-rater reliability were found between the different evaluation methods. Regardless of the evaluation methods, raters were consistent in their own evaluation, in their evaluation taken as a jury, and in their evaluation in comparison to other raters. This indicates the influence of the evaluation method on rater variability might be different in an L2 context. Of course, there were several shortcomings of this study. These will be briefly described, together with suggestions for future studies on writing evaluation. First, due to logistical difficulties, raters of this study were unable to communicate about the rating process beforehand. It is recommended ensure raters are in agreement about the formulation of the rating criteria beforehand, and fully understand any (analytic) rating rubrics. Second, this study used writing products from two student groups who were at opposite ends in terms of writing skill. Future studies might opt to use students from a continuum of possible skill levels to allow for a more equal division of writing product quality along the possible range. In addition to this, it is recommended to have students write using multiple different prompts to ensure internal testing validity. Third, only one reference text was used during the evaluation process. Future studies might opt to construct a rating scale instead of three texts to facilitate comparisons between the reference text(s) and the texts which are to be rated. Comparisons could also be made between using only one reference text and multiple reference texts to determine whether using multiple reference texts has a significant impact on the overall reliability of the evaluation method. Finally, this study used secondary school teachers with approximately two to five years of experience. A comparison between experienced and inexperienced teachers as raters might give some further insight into the effect of teaching experience on rating ability for text quality.

References

Archer, J. and McCarthy, B. (1988). Personal biases in student assessment. *Educational Research, 30* (2), 142-145.

Bachman, L.F., and Palmer, A.S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement*, *22*(1), 41-52.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, *19*, 309-316

Elving, K., & Bergh, H. van den (2016). Bruggen bouwen; Havisten leren om coherente teksten te schrijven. *Levende Talen Tijdschrift*, 17(1), 24-35.

Elving, K., & Bergh, H. van den (2017). Doen we weer Booster? Het effect van een digitale en interactieve schrijfcursus op tekstkwaliteit in havo 4. *Pedagogische Studiën*, 94(4), 330-347.

Feldt, L. S., & Kim, S. (2006). Testing the difference between two alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement*, *66*(4), 589-600.

Kox, F. & van den Bergh, H. (2018). An example of an L2-writing strategy: NOVSKEV. *Contribution to a special issue in honor of Gert Rijlaarsdam* Making Connections: Studies of Language and Literature Education. *L1-Educational Studies in Language and Literature, 18,* pp. 1-12. https://doi.org/ 10.17239/L1ESLL-2018.18.03.06

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, *12*, 115-117.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In: Kroll, B. (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 69–87). Cambridge University Press, Cambridge.

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement*, *69*(6), 887-912.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). Essay marking and the context problem. *Educational Research, 22*, 147-148

Huot, B.A. (1990). Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication 41,* 201-13.

Jacobs, H.J. et al. (1981). Testing ESL Composition: A Practical Approach. Rowley, Massachussets: Newbury House.

Llach, M. P. A. (2007). Lexical errors as writing quality predictors. *Studia linguistica*, *61*(1), 1-19.

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

Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. *Teaching of Psychology, 40*(3), 233-237.

McColly, W. (1970). What does educational research say about the judging of writing ability?. *The Journal of Educational Research*, *64*(4), 147-156.

Meuffels, B. (1994). *De Vergruisde Beoordelaar: opstellen over opstelbeoordeling* [doctoral dissertation]. Nijmegen: Faculteit der Letteren.

Nieva, V. F., & Gutek, B. A. (1980). Sex effects on evaluation. *Academy of Management Review, 5,* 267–276.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–256.

Oller, J.W., Perkins, K. (1980). *Research in Language Testing*. Massachusetts: Newbury House Publishers Incorporated.

Pilliner, A.E.G. (1969) Multiple Marking; Wiseman or Cox? *British Journal of Educational of Psychology, 39*, 313-315.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 108-122.

Rijlaarsdam, G., & Braaksma, M. (2004). De Smikkelcasus. Een praktijkvoorbeeld hoe leerlingen teksten leren schrijven zonder al te veel instructie (12-jarigen). *HSN 18: Het Schoolvak Nederlands*. p. 40-41.

Salkind, N. J. (Ed.). (2010). Spearman-Brown Prophyecy Formula. In *Encyclopedia of research design* (Vol. 3, pp. 1402-1404)). Thousand Oaks, California: SAGE Publishers.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33.

Spear, M. (1997). The influence of contrast effects upon teachers’ marks. *Educational Research, 39*, 229-233.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15* (1), 72-101.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2012). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*, *30*(1), 71-97.

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes* [Doctoral dissertation]. Utrecht: LOT

Van den Bergh, H., De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. *Measuring writing: Recent insights into theory, methodology and practices*, *27*, 23-32.

Weigle, S. (2002). *Assessing Writing* (Cambridge Language Assessment). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511732997

Wesdorp, H. (1981). *Evaluatietechnieken in het moedertaalonderwijs*. ’s-Gravenahge, Staatsuitgeverij.

Appendix A: Sources of variability for teaching ratings

Effect type	**Definition**	**Possible solutions as suggested by Wesdorp (1981)**
Signical effect	if multiple raters are asked to judge the same writing product without any further instruction, it is likely they will use varying approaches during evaluation. This is because there are several different ‘types’ of raters, each of which placing a varying degree of importance to specific writing product criteria.	**Improve transparency/clarity**: clear evaluation criteria, training of raters, discussions/meetings between raters
Halo-effect	external factors which are not related to performance might influence the judgement of raters. Irrelevant characteristics of either the writer (e.g. SES, student type) or the written product (e.g. handwriting) can be the cause.	**Remove irrelevant aspects of written product (if possible)**: students type out their texts, text are anonymised, independent raters are used
Sequential effect	raters are influenced by their ratings of preceding texts. This effect becomes more significant if there are more texts of similar quality before the current text. Is also influenced by the time of rating: judgement becomes stricter or more lenient due to ‘rater fatigue’.	**Minimise influence of order**: vary rating order, vary rating order per category or analytic rating, let raters use a rating scale with reference texts
Shifting norms	different raters have different standards of how frequently certain scores should be given. For instance, one rater might only give a perfect score for exceptional texts, while others might give perfect scores if the criteria of the assignment are met at the highest level.	**Limit variation in norms**: ask raters to strive for a certain average distribution of scores, let raters use a rating scale with reference texts, assign corrections after rating by evaluating personal averages and deviations

Appendix B: Writing prompt (in Dutch)

Adapted from: Rijlaarsdam & Braaksma (2004)

Tijd: 60 minuten

Lees de onderstaande informatie goed door. Je hebt dit nodig voor de opdracht.

Stel je voor:

Op de verpakking van de Engelse *Yum-Yum* chocoladerepen die je wel eens eet, heb je zien staan dat je twee gratis theatertickets kunt krijgen. Op de verpakking staat:

> **SAVE UP AND WIN TWO FREE TICKETS FOR THE LONDON THEATRE!!!**
>
> What you have to do:
> One entry token can be found on every wrapper of *Yum-Yum* chocolate bar you buy. Save up 10 of these tokens and send them in a stamped envelope to our address:
>
> Yum-Yum inc. Marketing Department
> New Headway Street 33
> 1023AB, London
> United Kingdom
>
> Please make sure to clearly state your name, address, postal code, and place of residence. The free (yes, that's right, FREE!) tickets will be sent your way as soon as possible. This promotional offer runs until June 26th, 2019.

Het is Mei 2019. Je hebt 8 punten bij elkaar gespaard, maar nu kun je nergens meer repen met punten vinden. Op de repen in de winkels zit geen spaarpunt meer, hoewel het nog geen 26 juni is. Toch wil je de twee theatertickets wel graag ontvangen. Je stuurt daarom 8 punten op en doet er twee hele wikkels zonder punt bij.

De opdracht:

Schrijf een brief die je meestuurt met de punten en de wikkels. Vertel waarom je geen tien punten kunt opsturen. Overtuig het bedrijf Smikkel ervan dat jij die twee theaterkaartjes wilt ontvangen en dat jij er niets aan kunt doen dat je geen tien punten hebt. **Zorg ervoor dat ze jou de theaterkaartjes toch sturen!**

Schrijf de brief in het Engels. Zorg dat je alle benodigde informatie volgens de advertentie erin hebt staan.

Appendix C: Test statistics for comparisons between evaluation methods.

Table 3.

Two-tailed t-test results per rater for comparisons between evaluation methods.

Evaluation methods	Teacher A			Teacher B			Teacher C			Teacher D		
	t	df	p	t	df	p	t	df	p	t	df	p
Holistic – Analytic	-5.9	19	<.01	-7.2	19	<.01	-1.9	19	.07	-7.9	19	<.01
Holistic – Relative	-3.2	19	<.01	-4.5	19	<.01	-1.6	19	.12	-4.2	19	<.01
Analytic – Relative	-1.3	19	.22	-.79	19	.44	-1.3	19	.21	-.2	19	.83