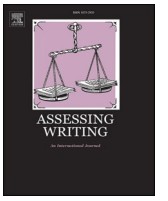
Assessing Writing 67 (2026) 100995

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



The relation between linguistic accuracy and scoring of Swedish EFL students' writing during a high-stakes exam

Christian Holmberg Sjöling *

Luleå University of Technology, Sweden

ARTICLE INFO

Keywords:
Writing assessment
Accuracy
High-stakes exams

ABSTRACT

This paper examines the effect of linguistic accuracy (e.g., the lack of form, grammatical, and lexical errors) on scoring during the high-stakes national test of English in Swedish upper secondary school. Teachers are expected to score their own students' texts with the help of assessment instructions containing benchmark texts (i.e., texts representing different score bands). The assessment instructions and the score bands provided to guide scoring are not explicit about how accuracy should influence scores. Two research questions were answered: As measured by ordinal regression, to what extent does linguistic accuracy predict rater scores? Do the texts scored by teachers reflect the graded example texts in terms of how linguistic accuracy predicts scores? The results revealed, amongst other things, that overall frequency of errors in texts significantly predicted scores as the model explained approximately 58 % of the variance in the outcome variable according to Nagelkerke's pseudo R-squared. Accuracy also had a similar effect on scores in texts rated by teachers as in the benchmark texts. In relation to the findings, it was concluded that accuracy may have more of an impact on scores than constructs that are more explicit components of the score bands such as lexical complexity.

1. Introduction and aim

Students' ability to express themselves with linguistic accuracy in writing is an important dimension of the English as Foreign Language (EFL) classroom and accuracy is a frequent component in summative and formative assessment (Pallotti, 2010). In summative assessment, several studies have found that more accurate texts are of higher quality than those that are less accurate (e.g., Choi & Cho, 2018; Cumming et al., 2006; Gebril & Plakans, 2013; Plakans et al., 2016). One reason for this is that linguistic accuracy is a common component of several scoring rubrics. For example, in the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT), the rubric states that texts awarded the highest score should contain "almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions", while texts awarded a lower score are characterised by "an accumulation of errors in sentence structure, word forms or use" (Educational Testing Service, 2023).

The study at hand concerns a similar high-stakes exam, namely, the national test of English (NTE) in Sweden. This test is conducted every year at the end of a course to ensure a fair and reliable assessment practice. It also serves as one of the few concrete reference points in terms of what the target language is at different levels of education for Swedish EFL teachers in upper secondary school. The test is high-stakes as the result on the test is tied to the students' final course grade, which is then used for university entrance – a higher score on the test will increase students' chance to access higher education where competition for admission is fierce, like law or

* Correspondence to: Luleå University of Technology, Department of Health, Education, and Technology, R101 971 87, Luleå, Sweden.
E-mail address: Christian.holmberg.sjoling@ltu.se.

https://doi.org/10.1016/j.asw.2025.100995
Received 21 December 2024; Received in revised form 21 November 2025; Accepted 23 November 2025
Available online 5 December 2025

medicine.

During the written part of the NTE, students are asked to complete a timed independent writing task. Then, teachers are provided with assessment instructions containing benchmark texts (i.e., texts representative of different score bands) by the Swedish National Agency for Education (SNAE) that should help teachers score texts written by their own students (e.g., University of Gothenburg, 2023a, 2023b). However, recent studies suggest that Swedish EFL teachers are not necessarily equipped to conduct such scoring (Bergström et al., 2022) and there appears to be a discrepancy between the expectations of the SNAE and scoring carried out by teachers during the NTE (Holmberg Sjöling, 2025).

In the present study, the aim is to examine the relation between accuracy (e.g., the lack of form, grammatical, and lexical errors) and scores, and whether or not teachers score in agreement with the assessment instructions during the NTE in Swedish upper secondary school. To do this, benchmark texts that are a part of the assessment instructions ($N = 72$ texts) have been compiled into a learner corpus, as have texts scored by teachers ($N = 93$ texts).

Accuracy is not an explicit part of the test's scoring rubric since the English subject in Swedish upper secondary school is characterised by a communicative approach to language teaching. The focus is instead on developing students' confidence and communication skills rather than productive knowledge of distinct linguistic features (see Section 2.3). However, whether accuracy has an effect on scoring in the aforementioned material has not previously received much attention, which is why the answer to the following research questions will be sought:

1) As measured by ordinal regression, to what extent does linguistic accuracy predict rater scores?
2) Do the texts scored by teachers reflect the benchmark texts in terms of how linguistic accuracy predicts scores?

The results of the study may have direct pedagogical implications for Swedish EFL teachers, curriculum developers, and test constructors. For example, the findings could indicate which accuracy features students in Swedish upper secondary school need to improve to increase the probability of receiving a higher score during the NTE. Outside of the Swedish context, the study may also contribute with additional knowledge of how accuracy relates to scoring of texts written by EFL learners, and the potential effect of accuracy during high-stakes exams.

2. Review of the literature

2.1. Accuracy and writing assessment

To set the stage for the present study, this section firstly explores accuracy as a construct and briefly considers it in relation to the learning context. Then, accuracy is situated within the Second Language Acquisition (SLA) research literature and, lastly, several relevant earlier studies regarding the relationship between accuracy and perception of writing quality in different writing contexts are presented.

The construct of accuracy can be defined in several ways. However, many researchers agree that it concerns learners' ability to produce language in line with norm-based correctness of form and structure (e.g., Foster & Wigglesworth, 2016; Plakans et al., 2016; Wolfe-Quintero et al., 1998). A language norm should, in this context, be understood as the accepted conventions and rules of language use within a particular setting, i.e., a set of rules which dictate expected grammatical structures and vocabulary. Accuracy, then, may be regarded as a measure of linguistic control over norm-based grammar and particular lexical errors, typically not including pragmatic or semantic accuracy (Plakans et al., 2016). Furthermore, when writers deviate from a language norm, it is traditionally referred to as an *error*, and written accuracy has commonly been studied by counting the number of these errors in a text (Plakans et al., 2016). This might come across as relatively straightforward, but identifying and labelling a linguistic occurrence as erroneous becomes challenging as the overall context and its relation to standard prescriptive target language norms and non-standard/non-native norms must be considered (Ellis, 2009; Pallotti, 2009; Polio, 1997). Thus, for a study on accuracy, it is essential to determine what the language norm is and what constitutes an error within that norm (see Section 3.2.1).

In the learning context, it may be contentious to consider language development in terms of errors. Ferris and Hedgecock (2023) question whether non-target-like-language produced by learners should even be considered *errors* or regarded as a natural consequence of evolving learner interlanguage (Corder, 1967). Nevertheless, they suggest that the error label may have practical implications for teachers since it can help them address rule-governed and patterned errors in the classroom through strategy training and instruction (Ferris & Hedgecock, 2023). It should be mentioned here that more advanced learners tend to commit errors that are in the more "shadowy areas" (Neff & Bunce, 2006, p. 698) of language use; i.e., these errors are less rule-governed, and instead concern infelicitous language use ("dispreferred forms") (Ellis & Barkhuizen, 2005, p. 59). In light of this, the results of the present study may help teachers and students understand which types of errors they must avoid in order to receive a high score during the NTE.

Accuracy can be placed within the wider framework of complexity-accuracy-fluency (CAF) in SLA research (Housen & Kuiken, 2009; Housen et al., 2012). This framework highlights three dimensions (complexity, accuracy, fluency) that are typically analysed to describe language proficiency, development, and quality in linguistic terms (Plakans et al., 2016). Firstly, complexity within CAF may be defined as "form-as-ambition" (corresponding to linguistic risk-taking) (Skehan & Foster, 2012, p. 200) and consists of several sub-constructs such as lexical (Bulté & Housen, 2012), and phraseological (Paquot, 2019) complexity. Pallotti (2009) regards complexity as the trickiest of the CAF constructs since it encompasses some eight aspects of language and communication (e.g., "lexical, interactional, propositional, and several types of grammatical complexity") (p. 593). The second construct of CAF, accuracy, concerns, as discussed, learners' ability to produce linguistic features in agreement with norm-based correctness in terms of form and

structure (Foster & Wigglesworth, 2016; Plakans et al., 2016; Wolfe-Quintero et al., 1998). It is regarded as a more internally coherent construct compared to, for example, complexity (Pallotti, 2009). The third construct of CAF, fluency, taps into how comfortably learners can retrieve and produce language in different contexts (Wolfe-Quintero et al., 1998). Fluency is often operationalised as length (e.g., text length) and perceived as a proxy for degree of automaticity (Ahmadian, 2011) since it, arguably, captures a learner's capacity to compose and produce language in the moment. Researchers have called for these constructs to be studied together (e.g., Larsen-Freeman, 2009; Plakans et al., 2016) in order to develop a more nuanced picture of how they interact with each other and develop over different levels of proficiency. Although the present study focuses on accuracy, it also takes into consideration an earlier study on the same material that investigated lexical complexity in order to develop a more nuanced understanding of how overall linguistic complexity affects grading in the material (Holmberg Sjöling, 2025; see below).

A number of studies have consistently shown a positive relationship between different forms of accuracy and perception of quality in various writing contexts. Choi and Cho (2018) studied how spelling errors influenced trained raters' holistic assessment of response quality by examining 148 responses across four L2 writing tasks. The participants went through online training which included scoring responses to different writing tasks in relation to a scoring rubric and benchmarks. Then, they were asked to score both original responses (with spelling errors) and corrected responses (without spelling errors) using the same holistic scoring rubrics. The findings showed that responses without spelling errors were scored more favourably than responses with spelling errors; and it was the quantity of errors that impacted scoring the most, not the type of errors, which may be a result of the participants' training and the scoring rubric (Choi & Cho, 2018).

Plakans et al. (2016), on the other hand, investigated the relation between scoring of 418 integrated writing responses from TOEFL and the linguistic features of CAF using correlations and regression. It was found that fluency was the strongest predictor of score variance while accuracy accounted for a smaller, albeit significant, part. Complexity had a significant effect on scores, but the smallest compared to all other variables. In terms of accuracy, morphological errors had largest effect on score variance compared to lexical and syntactic errors. The findings of Plakans et al. (2016) also reflect earlier work by Cumming et al. (2006) and Gebril and Plakans (2013), which studied grammatical accuracy across score levels in integrated writing tasks. In both studies, they conclude that texts with a higher score contain fewer errors while those with a lower score have more errors.

One study that is of particular interest to the present study is Dobrić (2024). They analysed 60 texts (approximately 27,000 words) written by advanced (expected to be between C1–C2) non-native learners taking their final English language proficiency test at university. The texts were rated independently by highly trained raters using a scoring rubric without explicit focus on errors, instead the focus was on successful language usage (Dobrić, 2024). In the study, it was hypothesized that errors affect raters even in assessment contexts where deliberate focus is placed on communicative language use. The hypothesis was tested by error tagging the rated texts. The results showed that errors explained a significant portion of score variability overall (around 50 % according to Nagelkerke's pseudo R^2) (Dobrić, 2024). Among other things, they conclude by suggesting that "it would be sensible to include reference […] in the relevant rating scales" to systematic error characteristics for specific assessment contexts (Dobrić, 2024, p. 12). This could increase the reliability and validity of the rating/grading process and be relevant in the training of raters (Dobrić, 2024). In the present study, raters are also asked to focus on successful language use over accuracy and Dobrić (2024) suggests that they may also give accuracy more weight than intended by the assessment instructions.

Taken together, the previous research shows that there is a clear positive relation between accuracy and assessment, i.e., texts with fewer errors are typically deemed as being of higher quality. Given the number of studies that support this relationship, it is to be expected that the relationship between accuracy and scoring will be similar in the current study as well.

Although the focus of this study is on accuracy, an earlier examination of the same material in this study analysed the effect of lexical complexity on scoring during the NTE (Holmberg Sjöling, 2025). It revealed that the benchmark texts and teachers considered lexical complexity when scoring, as a model of ordinal regression containing four predictors (lexical diversity, frequency and range of content words in relation to a reference corpus, and age of acquisition of content words) explained 33.4 % of the variance in the outcome variable (Nagelkerke's pseudo R^2). However, it also found a significant difference between the two types of raters, as one predictor of lexical complexity had a larger impact in scoring carried out by expert raters (Holmberg Sjöling, 2025). The identified discrepancy was discussed in relation to the nature of the English subject in Swedish upper secondary school as it relies on a communicative approach to language teaching, which is potentially problematic with regards to assessing linguistic complexity (i.e., there are no clear target features for any specific course) (see Section 2.3).

2.2. Errors by Swedish EFL learners in writing

Few studies have investigated the relation between Swedish EFL learners' written production and errors, especially in the context of a high-stakes exam. Köhlmyr (2003) conducted the only large-scale study of texts written during similar circumstances to the national test of English. She analysed grammatical errors in 383 written compositions by Swedish 16-year-old learners of English during the Swedish National Assessment programmes carried out in 1992 and 1995. Using contrastive analysis and error analysis, she identified 3525 errors in total and 75 % of them were related to three categories:

Some years after Köhlmyr (2003), Johansson and Geisler (2009) introduced the *Uppsala Learner English Corpus* (ULEC), which, at the time, consisted of essays (approximately 136,000 words) written by Swedish junior and senior high school students (age 14–19). ULEC is described as a resource that teacher students can utilise for degree projects and teaching purposes. Interestingly, teacher students who have analysed the corpus have identified several reoccurring errors that are somewhat on par with Köhlmyr's (2003) findings as many of them concern subject-verb agreement, article usage (generic reference), capitalisation (spelling), the s-genitive,

prepositions, and verb form (spelling, tense agreement) – most errors belong to the last category (Johansson & Geisler, 2009). It must be noted that the teacher students who have carried out analysis of ULEC may have consciously focused on specific error categories identified by Köhlmyr (2003) in their own studies (i.e., the material has not been exhaustively error annotated like in the present study; see Section 3.2.1).

2.3. The national test of English in Swedish upper secondary school

During their first year of English studies in Swedish upper secondary school, students are expected to complete the obligatory course English 5 (CEFR level B1.2).[1] The course itself is built on a communicative and functional approach and its focus is on language use and confidence rather than acquisition of specific linguistic aspects such as vocabulary and grammar (Erickson & Tholin, 2022). However, students are expected to learn how to produce texts characterised by "correctness […] variation and complexity", while also being linguistically adapted to different recipients, situations and purposes (Skolverket, 2022, p. 1).

The NTE is conducted each term to ensure that students have developed these abilities and to help teachers' assessment of their students' written productive ability (Erickson & Tholin, 2022). Also, the test is meant to ensure a fair and reliable assessment practice across the country and establish that students have successfully met the demands of the course they have studied (Erickson & Borger, & Olsson, 2022). Furthermore, the test is high-stakes as the result is tied to the grade students receive for the entire course, i.e., course grade should not deviate too much from test score; and these course grades are essential as they determine students' access to, for example, higher education, particularly programmes where admission is competitive, like law or medicine. It should be mentioned that the test consists of four parts including writing, speaking, and comprehension (reading and listening) tasks, but the focus of this paper is only on the written part.

Students are required to complete an independent writing task during the test in English 5. The task promotes topic familiarity by allowing students to choose relatively freely what to write about while also providing ideational and linguistic support, which is likely to increase students' chance to display their linguistic knowledge. Fig. 1 shows an excerpt of this writing task.

The texts written by students during the task are then typically scored by their teacher or other teachers at their school. Prior to carrying out scoring, assessment instructions are provided to teachers as they should get acquainted with benchmarks for the specific course and each score band. The assessment instructions state how teachers should carry out scoring by considering the focus of the communicative language approach, the demands of the specific course plan (see Skolverket, 2022), and different qualities of the texts such as purpose, content, genre, communicative strategies, structure, vocabulary, fluency, phraseology, idiomaticity, cohesive devices, grammatical structures, punctuation, and spelling (the latter two being the only potential indication of how accuracy should be considered).

To guide teachers' scoring, the SNAE provides benchmark texts representative of each score band, together with analytical comments provided by a group of scoring experts (Erickson & Borger, & Olsson, 2022). The analytical comments typically consist of various statements like: "The text contains a number of errors, but they do not impede comprehension" and are supported by textual examples of what that means. Together with the commentary material related to the English subject's syllabus, this is considered to be sufficient for teachers to carry out equal and fair assessment without specific training measures that are in place for other high-stakes examinations such as norming exercises or inter-rater reliability checks.

The scoring process is typically carried out by teachers at an individual school, either on their own, in pairs, or in larger groups, i.e., schools and teachers are given a lot of freedom to decide which option suits them best. Students' texts are assigned one out of six score bands: F (fail), E (pass), D, C, B, and A. For a text to be awarded E, C, or A, it must fully meet the knowledge requirements (descriptors) of each score band. However, if a text, for example, fulfils only some of the requirements for score band A, C, or E, it is assigned one of the lower score bands B, D, or F. The knowledge requirements are presented in Table 2.[2]

The knowledge requirements reflect how linguistic complexity should progress from lower to higher score bands. Different potential aspects of accuracy (e.g., spelling and punctuation) are mentioned in the assessment instructions as something that teachers should consider when scoring, but there is no explicit statement of what that difference looks like between different score bands (e.g., students should not commit any subject-verb agreement errors for the highest score band A).

3. Methodology

3.1. Materials

The learner corpus analysed in this study consists of two parts. The first part is comprised of benchmark texts, the benchmark texts that are part of the assessment instructions of the written part of the national test of English between 2012 and 2022. The other part of the material consists of authentic texts scored by teachers. These texts were written on the same topics and through the same period of time. Also, the later sample of the teacher scored texts were written digitally (without spell-checkers) and the earlier sample by hand. All of the teacher scored texts, and roughly half of the benchmark texts, are confidential, and therefore, linguistic examples of different

[1] These split-levels are used by the Council of Europe (see 2001, p. 33) and in this context indicate that the learners are on the upper end of B1 (*independent users*).

[2] The knowledge requirements are called 'grading criteria' as of 2021, but they are more or less the same in nature. From an international perspective, these should be understood as similar to scoring-band descriptors.

A Good Life

- What makes a good life? What is important to you?
The online magazine Outlook will focus on the subject A Good Life and you are now invited to write a text for the magazine.

You have **80 minutes** to write your text. It should be **at least 250 words** but not more than **600 words**. Make sure you have time to check what you have written.

Write a text about what is important in life to people in general – and to yourself. **Use some or all of the points on the next page** when planning and writing your text. You can also use the mind map for inspiration.

Explain – Give examples – Discuss – Compare

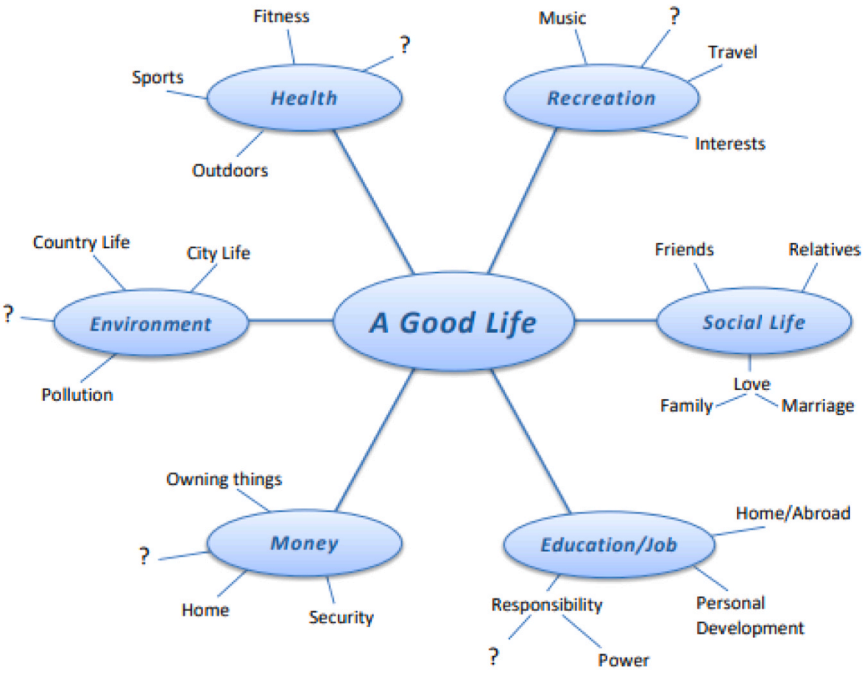


Fig. 1. Excerpt of writing instructions for a national test in English 5.
Any material from the national test of English included in this study is reproduced with permission from the Swedish National Agency for Education.

Table 1
Three prominent error categories identified by Köhlmyr (2003).

Category	Explanation	Example
Verb errors	Categorical issues of time and tense with regards to the simple present was the most prevalent type of verb error, i.e., it was either replaced by a different tense or construction in sentences where it was the expected correct form.	"*Usually my mother and I are going to Polen but I must work. [go]"* (Köhlmyr, 2003, p. 65).
Noun and article errors	Most noun errors were related to articles and categorical errors of number.	"*I am a glad and possitiv girl who loves the life, [life]*" (Köhlmyr, 2003, p. 52). "*My parent are divorst. [parents]*" (Köhlmyr, 2003, p. 42).
Concord	Defined by Köhlmyr (2003, p. 1 70) as errors related to "relations of number, gender, person, case, or tense". A large portion of these errors were related to subject-verb agreement.	"*I lives in a small town 22 Km north of X. [live]*" (Köhlmyr, 2003, p. 174).

Table 2
The knowledge requirements of English 5.

English 5		
E	C	A
In oral and written communications of various genres, students can express themselves in **relatively** varied ways, **relatively** clearly and **relatively** coherently.	In oral and written communications of various genres, students can express themselves in a way that is **relatively varied**, clear, coherent and **relatively structured**.	In oral and written communications of various genres, students can express themselves in ways that are varied, clear, coherent **and structured**.

phenomena are retrieved only from benchmark texts no longer under secrecy. In Table 3, the raw figures of the analysed material are presented.

The material consists of texts written on eight different topics such as the one outlined in Fig. 1 (see Section 2.3), but the task itself is the same (i.e., a timed independent writing task with ideational and linguistic support). Despite the topics being different, the intention is for them to test the same knowledge requirements of the particular course (see Section 2.3). The texts are analysed across topics, which might not be ideal from a linguistic perspective, but it was deemed necessary due to the limited number of benchmark texts available for each topic (approximately 8–12 per topic). This approach is supported by how the test constructors sometimes change topic, but maintain the same benchmark texts for the new topic.

3.2. Procedure

3.2.1. Error annotation

The first step of the analysis was to carry out manual error annotation of the raw texts with the *Université Catholique de Louvain Error Editor* (UCLEEv2) (Granger et al., 2023). UCLEEv2 was developed to speed up manual error annotation and this study utilises a tagset called the *Louvain Error Tagging Manual, version 2* (Granger et al., 2022) (for an early account of different error tagging systems in Learner Corpus Research, see Diaz-Negrillo & Dominguez, 2006).

The tagset used in this study consists of seven error domains and a total of 54 specific tags accompanied by several guidelines and principles to strengthen the reliability of the annotation process (Granger et al., 2022). Table 4 contains a short description of the seven errors domains (for further explanation, see Granger et al., 2022, pp. 9–34).

The reason for the creation of the infelicities category was to improve the reliability of the other categories as it covers errors that are unclear, stylistically awkward, or inappropriate, but not clear lexical or grammatical errors. In other words, it is the most subjective error category reliant on the judgement of the annotator (Granger et al., 2022).

Error annotation is a complex endeavour and dependent on the context the analysed material was derived from. In this study, an error is broadly defined as a linguistic occurrence, which is not acceptable in the specific assessment context and in relation to the error taxonomy of the *Louvain Error Tagging Manual, version 2* (Granger et al., 2022). The assessment context and the tagging manual in this scenario rely on a standard English (American/British) norm, i.e., linguistic occurrences that, for example, do not adhere to general rules of English grammar, spelling or lexical restrictions are tagged as erroneous.

The initial tagging occurred in two steps: first, the author tagged a text followed by a short break and, then, the tagged text was checked once again to establish that all tags were accurate and that no errors were missed. When this process was completed for every individual text, every tag out of the 54 possible that occurred was checked internally to eliminate any potential errors. For example, there were 320 article errors in the material tagged <GA> . In UCLEEv2, each of these occurrences were investigated to make sure that they actually were article errors, and if not, the tag was changed to the correct one. Then, to prepare the material for the next step described in the next Section (3.2.2), the material was "cleaned" (for example, spelling and grammar errors were corrected) in relation to the inserted tags (the tags themselves were also removed after the correction was carried out). During this step, any wrongly tagged errors, or errors without a tag, were also addressed. Lastly, approximately 10 % of the material (18 texts) was checked for intra-rater reliability almost six months after the initial tagging. This process included inspection of 819 tags, and only four of them (0.49 %) were deemed to be tagged inaccurately. In these 18 texts, there were 21 instances where an error occurred (primarily spelling errors), but no tag was inserted. The intra-rater reliability check suggests a high level of accuracy for the assigned tags while it also indicates that the raw frequencies of error tags in both sets of texts presented in Table 6 are likely slightly higher.

A coding error can occur when the annotator fails to insert an error tag where there is an actual error (or vice versa), which affects the reliability of the error tagging. This can be caused by annotation fatigue or when the linguistic occurrence that is thought of as an error is in the more "shadowy areas" of language use (Neff & Bunce, 2006, p. 698). One such example is: "Uganda, I visited 5 years ago when a storm had raged […]", which was deemed to be a word order error since it deviates from the subject-verb-object pattern of English (as it instead is object-subject-verb). This example was tagged as an error in relation to the rest of the text since it caused confusion rather than, for example, being a functioning stylistic device. It is possible that more clear-cut errors which contravene general rules of English (e.g., form and grammar errors) have a higher degree of annotation reliability as they are easier to spot while, for example, lexical and lexico-grammar errors may be less reliable. This is, however, an issue that almost all studies concerned with accuracy are affected by, and despite issues with reliability and subjectivity, error annotation of "a learner corpus allows us, at least and at most, to systematize our intuitions" (Milton & Chowdhury, 1994, p. 129).

Table 3
Overview of texts in English 5.

Benchmark texts							
Score band	**F**	**E**	**D**	**C**	**B**	**A**	
N	9	17	12	15	8	11	**72**
Mean length	285	308	378	398	492	448	
Words tot.	2498 (9.25 %)	5182 (19.19 %)	4601 (17.04 %)	5988 (22.18 %)	3822 (14.16 %)	4908 (18.18 %)	**26,999 (100 %)**
Teacher scored texts							
Score band	**F**	**E**	**D**	**C**	**B**	**A**	
N	14	16	16	16	16	15	**93**
Mean length	334	338	392	442	471	489	
Words tot.	4678 (12.35 %)	5301 (14 %)	6198 (16.37 %)	6923 (18.3 %)	7478 (19.74 %)	7301 (19.28 %)	**37,879 (100 %)**

Table 4
The Louvain error tagging system.

Definition	**Description**	**Example**
Form	Spelling errors and morphological errors on derivational affixes (including homophones)	Errors related to misuse or omission of capital letters: "english" instead of "English"
Grammar	Errors that contravene the general rules of English grammar	Errors related to misuse of indefinite determiners: "There are too much students…" instead of "There are too many students…"
Lexis	Errors involving the conceptual, collocational or connotative properties of words or phrases	Lexical errors affecting adjectives: "Resorting to violence might seem like a comprehensible reaction" instead of "Resorting to violence might seem like a understandable reaction"
Lexico-grammar	Errors where the lexico-grammatical properties of a word have been violated, i.e., erroneous dependent prepositions, erroneous complementation patterns, or countable/uncountable noun confusion	Errors concerning adjectives used with the wrong dependent preposition: "representative for their own identities" instead of "representative of their own identities"
Word redundant/ missing/order	Unnecessary use of words, missing necessary words, or misordered words	The inclusion of unnecessary words: "I only want to to try and do my best" instead of "I only want to try and do my best"
Punctuation	Errors that target punctuation problems, e.g., confusion between punctuation markers, missing or redundant markers	Lack of commas where appropriate: "However it was never her fault." instead of "However, it was never her fault."
Infelicities	Issues concerned with register, questions of political correctness, stylistic issues, and unclear sequences of words or sentences	Unclear sequences of words/sentences: "To have the it of was on yesterday very nice."

3.2.2. Potential occasion analysis

Potential occasion analysis brings together the set of manually error annotated data (as described in Section 3.2.1) with POS-tagging of the cleaned version of the texts. This allows one to calculate errors in relation to relevant word classes, e.g., modal verb errors in the tagged data out of the total number of modal verbs in the individual text (Thewissen, 2021). Table 5 shows examples of potential occasion analysis formulae.

The clean version of each text was POS-tagged with CLAWS7.[3] The tags were then counted with an R-script to establish the value of the denominator (the dataset is available through the Open-Science Foundation, https://osf.io/user/htbxs). Then, the analysis was carried out by dividing the tally/count of the error tag by the total count of this same POS category. By doing this, one "puts findings into perspective by considering the ratio of accurate vs. inaccurate use" (Thewissen, 2021, p. 312) while also accounting for text length. Not taking text length into account can lead to misinterpretation of the findings since longer texts can contain more errors, but at the same time be more accurate. After the potential occasion analysis, the output was statistically analysed.

3.2.3. Statistical analysis

The data was initially plotted with box plots and visually inspected. Then, ordinal regression analysis with interactions was employed, as the model is suitable when there is an ordinal dependent variable (i.e., when there is a clear order or ranking between different categories like score bands F to A). This model shows the influence of the independent variables (predictors, specifically types of errors) on the dependent variable (score bands), while also accounting for any potential differences in terms of the influence of the predictors in the two sets of texts. In order to use certain statistical models, specific assumptions have to be met (e.g., for ordinal regression there should be no multicollinearity, and the proportional odds/parallel lines assumption should hold) (Laerd Statistics, 2024). All assumptions were checked and met. The following packages in *R* (R Development Core Team, 2023) were used to carry out the analysis: ordinal (Christensen, 2023), car (Fox et al., 2012), AER (Kleiber et al., 2020), rcompanion (Mangiafico, 2024), MASS (Venables & Ripley, 2002), and brant (Schlegel et al., 2018).

[3] The "FREE CLAWS web tagger" available at Free CLAWS web tagger (lancaster.ac.uk) was utillised for potential occasion analysis since it has shown high accuracy for tagging learner data (around 95 %) (Granger et al., 2009).

Table 5
Examples of potential occasion analysis formulae.

Example 1	Example 2
Nominator = Modal verb errors in the error-tagged text **POS denominator** = Overall modals in POS-tagged text	**Nominator** = Personal pronoun errors in the error-tagged text **POS denominator** = Overall personal pronouns in POS-tagged text

Note. POS = part of speech. Reproduced from Thewissen (2013, p. 82)
Reproduced from Thewissen (2013, p.

4. Results

4.1. Raw frequencies of error tags in both sets of texts

Table 6 displays a summary of the raw frequencies of error tags in the benchmark texts and the teacher scored texts after the manual annotation process was completed. The error categories are sorted by frequency as are the error tags within each category.

Table 6 shows that 43 out of 54 possible tags available in the *Louvain Error Tagging Manual, version 2* (Granger et al., 2022) occurred in both sets of texts. In the benchmark texts and texts scored by teachers, spelling errors are the most frequent type of error due to issues with, for example, capitalisation (1), doubling of consonants/vowels (2), and homophones (3):[4]

(1) english, south Africa, sweden
(2) allways, manny, litterally
(3) confusion of their/there and it's/its

The second most frequent type of errors are punctuation errors, specifically, missing punctuation markers after introductory elements (4), and absence in relation to dependent and independent clauses (5):

(4) "For example people who listen to rap […]"
(5) "If the writer is happy the song will be happy […]"

The third most frequent type of error concerns all errors of agreement between a subject and a verb. For example, use of a plural verb when there is a singular subject (6) or vice versa (7):

(6) "My grandmother like electronics, she likes her phone […]"
(7) "These children often believes that their future […]"

These findings may be regarded as on a par with errors committed by Swedish learners of English identified by Köhlmyr (2003) and Johansson and Geisler (2009). One category, however, which has not been explored much earlier, stands out in this investigation, namely, the category concerning words that are either missing (8), redundant (9) or in the wrong order (10) (Word R/M/O):

(8) "I am a 19 year old newly student from Sweden […]"
(9) "Am I am going to cook or […]"
(10) "Uganda, I visited 5 years ago when a storm had raged […]" (object-subject-verb instead of the typical subject-verb-object pattern of English; was deemed an error in relation to the rest of the text as it caused confusion rather than, for example, being a stylistic device)

This might be caused by the texts being written during a high-stakes exam and, as such, learners may forget to include necessary words or include unnecessary words due to the pressure of the situation. Some tags in the error taxonomy did not occur in the tagging process as there were no clear errors related to single logical connectors (e.g., using "Really," instead of "Actually," in the beginning of a sentence) or complex logical connectors (e.g., using "As a conclusion," in the beginning of a sentence instead of "To conclude,"). The absence of these types of errors is likely due to the formality of the texts, which is not high and, as such, does not warrant use of, for example, complex logical connectors.

4.2. Initial exploration of overall accuracy in relation to score bands and text length

In the NTE, the assessment instructions outline how assessors should take many textual dimensions into consideration like grammar, vocabulary, and content (see Section 2.3). Accuracy is also mentioned as a construct to consider. However, the instructions make it clear that texts may contain errors as long as they are not too "distorting". Similar to how many dimensions must be considered when scoring a text, it is important to consider the overall accuracy of a text and not just individual errors. Thus, for each individual

[4] The results section contain examples of inaccurate language use without corrections as the examples are deemed mostly straightforward enough.

Table 6
Raw frequencies of errors tags in both sets of texts.

Type of error	Benchmark	Teacher
Form	1305 (42.97 %)	1518 (42.1 %)
Spelling errors	1291	1514
Morphological errors on derivational affixes	14	4
Grammar	702 (23.12 %)	894 (24.79 %)
Errors of agreement between a subject and its verb	193	282
Article errors	147	173
Errors of misuse of tense or aspect	96	110
Errors affecting noun number	84	110
Errors affecting noun case	32	42
Errors due to confusion between word classes	19	48
Adverbial order errors	26	28
Errors affecting personal pronouns	23	19
Errors affecting primary/modal/semi- auxiliaries	6	16
Errors affecting indefinite determiners	8	13
Errors affecting demonstrative determiners	7	13
Errors affecting demonstrative pronouns	12	7
Errors involving non-finite/finite verb forms	16	2
Errors affecting possessive pronouns	8	9
Errors affecting relative/interrogative pronouns	4	11
Unclear pronominal reference	7	4
Errors affecting comparatives/superlatives	4	3
Errors affecting reflexive/reciprocal pronouns	3	3
Errors affecting verb morphology	4	0
Adjective order errors	2	0
Errors affecting indefinite pronouns	1	1
Punctuation	411 (13.54 %)	575 (15.94 %)
Missing punctuation markers	366	566
Confusion of punctuation markers	25	2
Redundant punctuation markers	20	7
Word R/M/O	284 (9.35 %)	318 (8.82 %)
Word redundant	109	152
Word missing	112	133
Word order	63	33
Lexis	200 (6.59 %)	152 (4.21 %)
Lexical errors affecting verbs	69	79
Lexical errors affecting nouns	42	33
Lexical errors affecting adjectives	20	15
Lexical errors affecting prepositions	30	2
Errors affecting fixed word combinations	15	5
Errors involving coordinating conjunctions	12	6
Lexical errors affecting adverbs	7	9
Errors involving subordinating conjunctions	5	3
Lexico-grammar	79 (2.6 %)	105 (2.91 %)
Nouns used with the wrong dependent preposition	59	53
Verbs used with erroneous/missing/redundant preposition	11	29
Errors involving countable/uncountable nouns	5	16
Adjectives used with the wrong dependent preposition	4	7
Infelicities	56 (1.84 %)	44 (1.22 %)
Total number of errors	3037 (100 %)	3607 (100 %)

text, the total number of errors have been divided with the number of tokens (to take text length out of the equation). The final value for each text has been plotted in Fig. 2 in relation to score bands and whether or not the text was a benchmark text or a text scored by a teacher.

Fig. 2 shows a clear gradual progression of error rate (overall number of errors) between the lowest score band F and the highest score band A in both the benchmark texts and texts scored by teachers (i.e., texts receiving a lower score are less accurate than those receiving a higher score). Additionally, there is much overlap between the higher score bands C, B, and A, which might be because raters regard a small number of errors as acceptable while other dimensions like content or cohesion are more important. Overall, the findings are in line with earlier research also showing a clear relation between accuracy and perceived quality of a text (Choi & Cho, 2018; Cumming et al., 2006; Gebril & Plakans, 2013; Plakans et al., 2016).

4.3. Results of ordinal regression models for selected predictors

A statistical model called ordinal regression was employed to determine the effect of an independent variable (overall accuracy; a predictor) on the dependent variable (different score bands; F to A) (see Section 3.2.3).

In *R*, the equation of the first ordinal regression model was: Score bands ~ Text *(Errors). The Text variable is categorical and refers

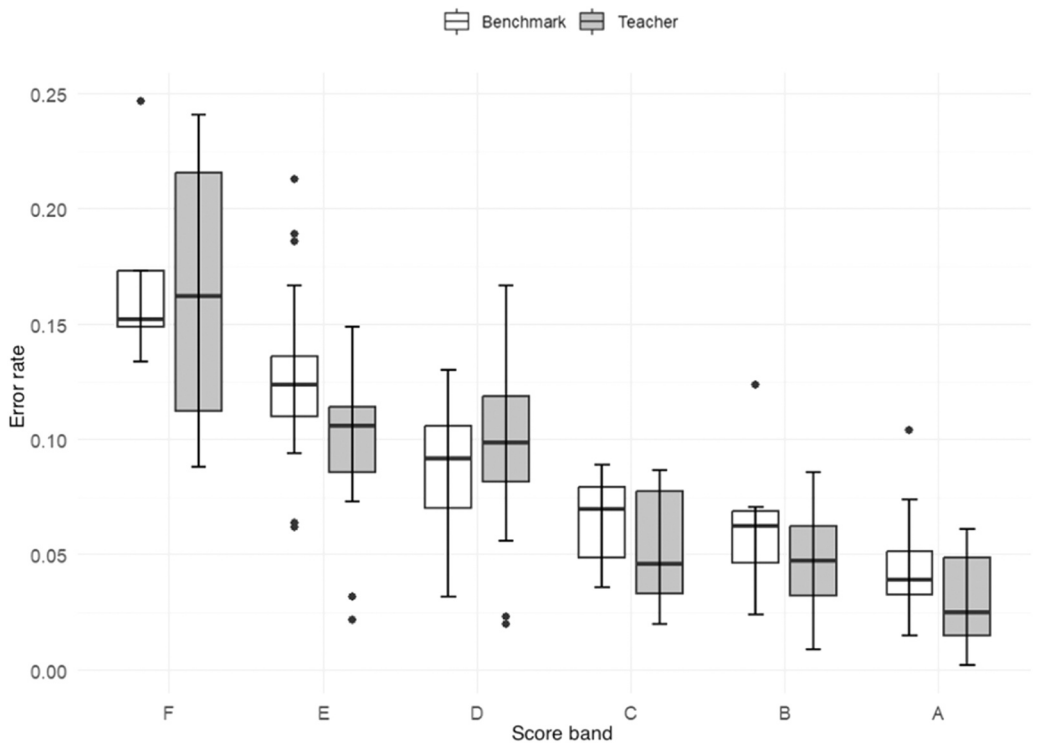


Fig. 2. The overall accuracy of individual texts in relation to score bands.

to if texts were scored by experts (i.e., the benchmark texts) or teachers while the * symbol in *R* indicates that the model includes both the main effect and interaction between Text and the variable in the parentheses (Errors). The interaction effect shows how one variable (Errors) impacts the relationship between another variable (Text) and the dependent variable (Score bands), which allows one to determine if errors have a stronger or weaker effect on score bands in the benchmark texts or in texts scored by teachers. The independent variable (Errors) is the same variable that was plotted in Section 4.2 (i.e., the value of the variable is the number of errors in a text divided by the number of tokens of the text). The fit of the ordinal regression model was evaluated by comparing a model with the predictor (Errors) to a null model (i.e., a model without predictors). As indicated by ANOVA, the ordinal regression model for the benchmark texts and the teacher scored texts (AIC = 466.45) had a significantly better fit than the null model (AIC = 598.48; x^2 = 138.03, df = 3, p = <0.001). Further assessment of the model's fit was done with Nagelkerke's pseudo R^2, which showed that the model explains approximately 58 % of the variance in the outcome variable. The model is summarised and presented in Table 7.

The model shows that overall accuracy (Errors) has a significant effect on score bands in the benchmark texts and in texts scored by teachers. Furthermore, there appears to be no significant difference in terms of the effect of accuracy on scoring in the benchmark texts and texts scored by teachers (i.e., there is no interaction effect), which is positive as it suggests that errors are considered similarly in relation to score bands by both groups of raters. In Holmberg Sjöling (2025), analysis of lexical complexity in the same material using the same statistical modelling revealed that lexical complexity had a stronger effect on score bands in the benchmark texts than in the teacher scored texts. A similar discrepancy does not appear to exist in terms of accuracy, which is further discussed in Section 5. The

Table 7
Summary of ordinal regression model with one predictor.

Coefficients					
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**95 % C.I.**
TextTeacher	0.14	0.64	0.22	0.83	-1.11 – 1.39
Errors	40.85	5.81	7.03	< 0.001	29.90 – 52.73
TextTeacher:Errors	3.99	6.75	0.59	0.56	-9.28 – 17.27
Threshold coefficients					
	Estimate	**Std. Error**	**z value**		
A | B	1.01	0.53	1.91		
B | C	2.19	0.53	4.13		
C | D	3.55	0.57	6.21		
D | E	4.91	0.64	7.63		
E | F	6.93	0.76	9.08		

significant predictor Errors is explored additionally through an effect plot presented in Fig. 3.

Fig. 3 illustrates how the probability of receiving the highest score A increases when the total number of errors in a text decreases. An opposite relation is found for the lowest score F as the probability of a text being assigned the lowest score F increases when the total number of errors in a text increases. The curvature of the slopes for scores E, D, C, and B suggests that the probability of receiving a lower score increases as error rate increases, but this appears to level off for higher scores in a gradual way (as illustrated by how the peak of the slope moves to the left for higher scores). This can be seen as positive since it indicates that there is a gradual consideration of accuracy in relation to scores despite not being mentioned as an explicit grading criterion (see Section 2.3). This finding is further discussed in Section 5.

One issue with the plotted predictor (Errors) is that it conflates all individual error tags into one, which may be less ideal from a linguistic perspective as it comes at the expense of possible linguistic description. Therefore, four error tags were selected for further analysis based on their raw frequency of occurrence (i.e., they represented a significant portion of the total errors) and distribution of these occurrences across different score bands in order to meet the assumption of proportional odds/parallel lines in ordinal regression (i.e., the effect of a predictor must be constant across all score band levels). Prior to statistical analysis, potential occasion analysis (see Section 3.2.2) was carried out and the nominator-denominator pairings of the selected tags are presented in Table 8.

The values produced by dividing the numerators with the denominators were analysed in an additional model of ordinal regression (see Section 3.2.3). The equation of the model in *R* this time was: Score bands ~ Text *(FS + GA + QM + GVT). Firstly, to assess model fit, a null model without predictors was compared to the model with predictors. The model with predictors (AIC = 479.48) was a significantly better fit than the null model (AIC = 598.48; $x^2 = 137$, df = 9, $p = <0.001$). Nagelkerke's pseudo R^2 was also used to assess model fit and it showed that the model explains approximately 58 % of the variance in the outcome variable. The model is summarised in Table 9.

The model summary shows that three predictors have a significant impact on score bands: spelling errors, article errors, and misuse of tense or aspect. There is no significant difference (i.e., no interaction) for the effect of these predictors on score in the benchmark texts and texts scored by teachers, which can be considered positive as it indicates that teachers are able to transfer the expectations of the assessment instructions to their own practice. One potential explanation for this is that these errors are relatively straightforward to identify (i.e., there are clear rules for what constitutes errors related to spelling, articles, and tense/aspect). This finding is further discussed in Section 5. The significant predictors are explored through individual effect plots. The non-significant predictor related to missing punctuation markers has also been plotted in Fig. 7. Spelling errors (FS) were the strongest predictor of score band and is plotted in Fig. 4.

The shape of the slopes shows that the probability of receiving a higher score (y-axis) is affected by the number of spelling errors a text contains (x-axis). For example, there is a clear opposite relation between the slopes for scores A and F, which illustrates how texts awarded the highest score have a higher probability of receiving the highest score if they contain fewer errors while those awarded the lowest score have a higher probability to receive the lowest score if they contain more errors. These spelling errors include issues with capitalisation (11), doubling of consonants/vowels (12), misordering/substitution of letters (13), homophones (14), inaccurate use/omission of blanks/hyphens between parts of words or words (15), inclusion of mother-tongue words (which are also covered in this

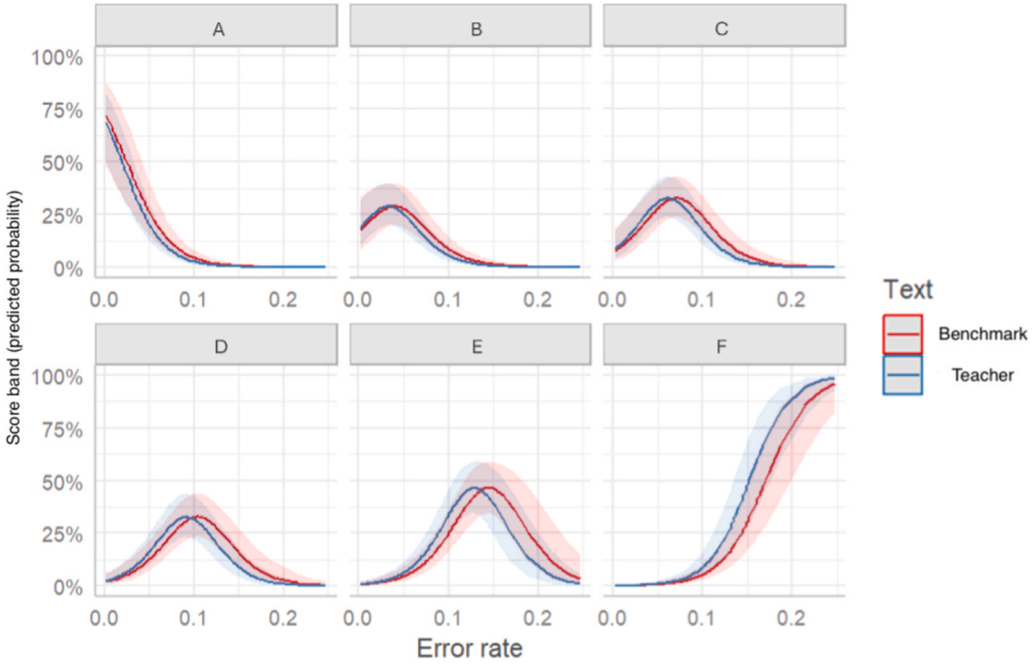


Fig. 3. The effect of errors on scores.

Table 8
Potential occasion analysis: error tag-denominator pairings.

Error type	Numerator (error tag)	Denominator[a]
Form spelling error	<FS>	Total tokens
Grammar article errors	<GA>	NOUNall
Grammar verb tense error	<GVT>	Vlex
Missing punctuation marker	<QM>	PUNC

[a] The denominator categories are adapted from Thewissen (2015) and primarily consist of several different POS-tags from CLAWS7. **NOUNall**=ND1, NN, NN1, NN2, NNA, NNB, NNO, NNO2, NNT1, NNT2, NNU, NNU1, NNU2, NNL1, NNL2, NP, NP1, NP2, NPD1, NPD2, NPM1, and MPM2. **Vlex**=VV0, VVD, VVG, VVGK, VVI, VVN, VVNK, and VVZ. **PUNC**=all punctuation markers (explanation of the individual tags are available at UCREL CLAWS7 Tagset).

Table 9
Summary of ordinal regression model with four predictors.

Coefficients					
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**95 % C.I.**
TextTeacher	0.18	0.61	0.29	0.77	-1.03 – 1.39
FS	44.78	9.93	4.51	< 0.001	25.92 – 65.11
GA	29.39	6.48	4.53	< 0.001	17.25 – 42.82
QM	-0.72	2.44	-0.3	0.77	-5.58 – 3.98
GVT	18.17	5.2	3.49	< 0.001	8.48 – 29.09
TextTeacher:FS	-6.35	12.09	-0.52	0.6	-30.18 – 17.56
TextTeacher:GA	1.93	11.19	0.17	0.86	-20.23 – 23.8
TextTeacher:QM	2.94	3.17	0.93	0.35	-3.22 – 9.24
TextTeacher:GVT	2.74	7.75	0.35	0.72	-12.62 – 17.9
Threshold coefficients					
	Estimate	**Std. Error**	**z value**		
A | B	0.36	0.5	0.73		
B | C	1.51	0.49	3.08		
C | D	2.81	0.52	5.42		
D | E	4.18	0.58	7.16		
E | F	6.34	0.72	8.8		

category), and slips of the pen/accidental keyboard strokes that cause spelling errors. The latter form of errors is theoretically possible, but it is difficult to establish whether they have occurred because of, for example, stress due to the high-stakes context (actual slips of the pen) or lack of word knowledge.

(11) american, african, wow (referring to the computer game World of Warcraft)
(12) bodys, lifes, Knowels (referring to pop singer Beyoncé Knowles)
(13) there/their/they're, and to/too/two
(14) baseballbats, computermade, ingame
(15) och (Swedish conjunction equivalent of "and"), mobbad (Swedish word for the verb "bullied"), Syrien (the Swedish word for the proper noun "Syria")

Spelling errors were the strongest predictor. The second strongest predictor, article errors (GA), is plotted in Fig. 5.

The effect plot shows that texts with a higher frequency of article errors tend to have a higher probability of being assigned one of the lower score bands F, E, or D. Then, the slopes indicate how texts with no article errors have the highest probability of being assigned to score bands C, B, or A, which is contradictory in a sense as, for example, A is the highest possible score band. This category includes all erroneous article use which does not adhere to the general rules of English grammar, such as issues with the definite article (16), indefinite articles (17), and omission of definite/indefinite articles (18):

(16) "The life is an event"
(17) "of a drug or a overdose"
(18) "so I can be useful in __ youth centre"

The third significant predictor of score in the model of ordinal regression was misuse of tense or aspect (GVT), and this has been plotted in Fig. 6.

Similar to the two other significant predictors, the effect plot shows that the probability of a text being assigned score band F, E, or D increases when tense/aspect errors occur, while texts with none of these errors have the highest probability of being assigned to score band C, B, and A. These errors include present continuous for simple (19), present simple for present perfect (20), simple past for present perfect (21), present perfect for simple past (22), and present simple for future (23).

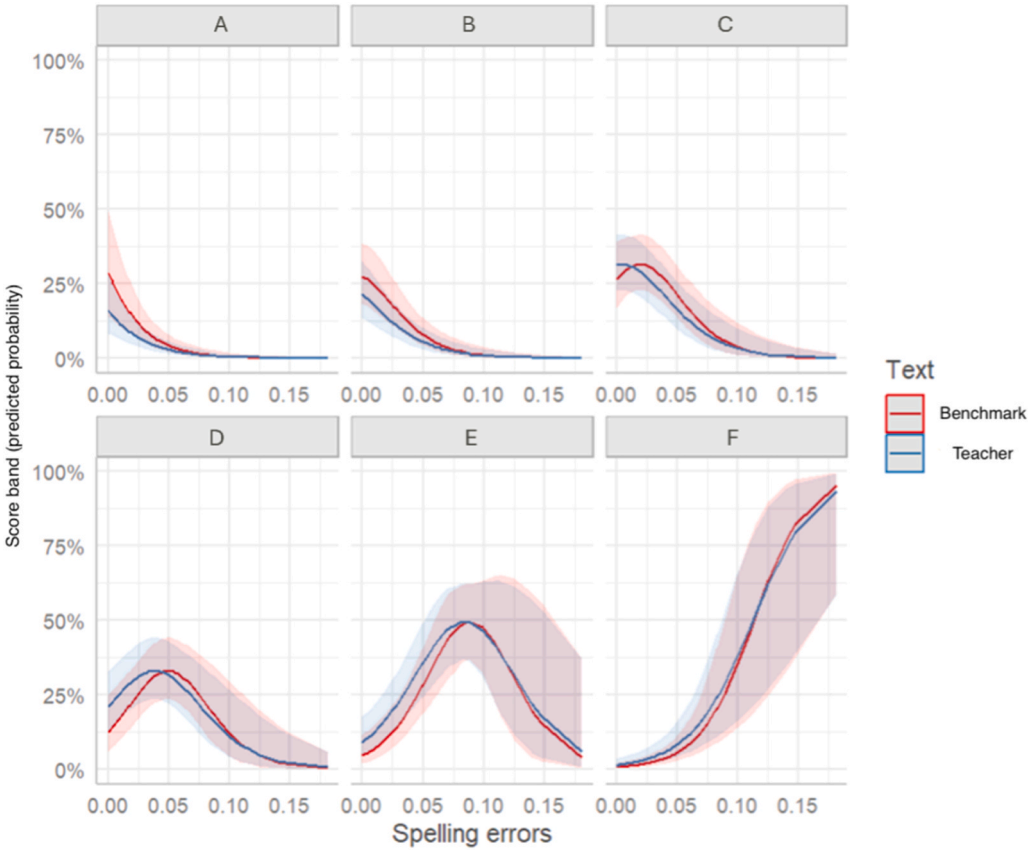


Fig. 4. The effect of spelling errors on score bands.

(19) "I am drinking coffee every morning" (drink)
(20) "Since I was a little child my dream was to make the world a better place" (has been)
(21) "They won several awards so far" (have won)
(22) "and at home the radio often has been on" (was)
(23) "She helps you with the project later" (will help)

The final predictor included in the model, QM (missing punctuation markers), had a non-significant impact on score bands, but has nevertheless been plotted in Fig. 7 for further exploration.

Fig. 7 shows that errors consisting of missing punctuation markers occur relatively frequently across all score bands. The reason why there appears to be a small cline from low to high is because the texts assigned a lower score band are simpler and, hence, do not present as many opportunities where, for example, commas should be used. Most of the errors are related to lack of commas before conjunctions in compound sentences (24), after introductory elements (25), absence with regards to dependent and independent clauses/phrases (26), and absence when separating things in a list (27).

(24) "People have to make choices everyday but the most important things"
(25) "However I know people get affected"
(26) "If I get the opportunity to volunteer in Chembapur I know what to expect"
(27) "She enjoys reading, writing drawing and painting"

Essentially, this suggests that teaching Swedish students correct usage of English punctuation markers is not a goal in the first course of English in Swedish upper secondary school (this finding is discussed together with communicative language teaching in Section 5).

5. Discussion and conclusion

The study at hand revealed that accuracy has a significant impact on scoring during the high-stakes national test of English in Swedish upper secondary school and the identified errors largely reflect those identified earlier as typical of Swedish learners in similar

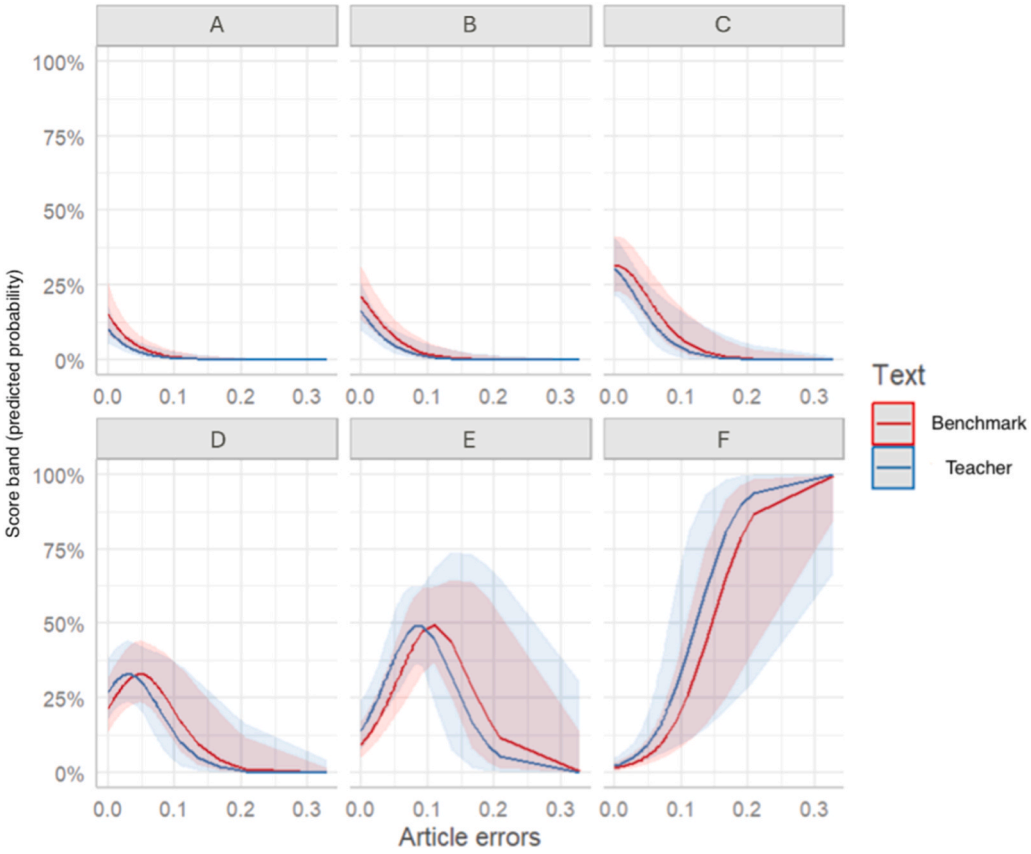


Fig. 5. The effect of article errors on score bands.

contexts (e.g., Johansson & Geisler, 2009; Köhlmyr, 2003). The findings also suggest that teachers consider accuracy to the same extent as the benchmark texts that have been scored by experts, which is positive since it indicates that teachers are able to carry out scoring with regards to accuracy in line with the expectations of the assessment instructions (see Section 2.3). In fact, for both sets of texts, overall accuracy explained a large portion (approximately 58 %), of the variance in the outcome variable according to Nagelkerke's pseudo R^2 (see Section 4.3). This is in line with Dobrić (2024) who found that errors accounted for a significant portion of the variability of the rating awarded to texts in their material (approximately 50 % also according to Nagelkerke's pseudo R^2). The rating process in Dobrić (2024) was not unlike the assessment practice in this study since their raters did not receive a rubric with an explicit focus on different forms of errors either. Instead, their assessment instructions had a focus on successful language use (Dobrić, 2024), which is similar to the assessment instructions of the national test of English (see Section 2.3). When synthesized, it appears as if the accuracy of a text has a significant impact on the score band it is assigned to by an assessor even when the assessment instructions do not include it as an explicit component. This is also interesting in light of Holmberg Sjöling (2025) as lexical complexity, despite being a more explicit component of the assessment instructions, explained a smaller portion of the variance in the outcome variable (33.4 %) according to Nagelkerke's pseudo R^2. The reason for this difference might be because accuracy, particularly spelling errors, are easier to identify compared to lexical complexity.

The findings signal a need for improvement of the design and theoretical foundation of the English subject in Swedish upper secondary school and call into question the validity argument for the national test of English. Students are expected to learn how to write texts characterised by "correctness […] variation and complexity" (Skolverket, 2022: 1), but at the same time, the focus of the course plans is on confidence and language use over acquisition of distinct linguistic features (Erickson & Tholin, 2022). The assessment instructions, on the other hand, mention that assessors should consider spelling and punctuation, but not in a clear and explicit way (see Section 2.3). In other words, the qualitative statements of the course plans and assessment instructions are at odds with what the benchmark texts contain empirically, which disadvantage teachers and students (i.e., the test does not necessarily test for features students are being taught). A clear example of this paradoxical situation is the non-significant predictor QM (missing punctuation markers) plotted in 8 (see Section 4.3). The course plan of English 5 (see Section 2.3) does not clearly state that students should learn correct punctuation, but in the assessment instructions, it is mentioned that punctuation should be taken into account (to some degree, it is not clear how much). However, both the benchmark texts and the teacher scored texts have a relatively high frequency of errors related to missing punctuation markers distributed somewhat evenly across all different scores (which is why it is a non-significant predictor).

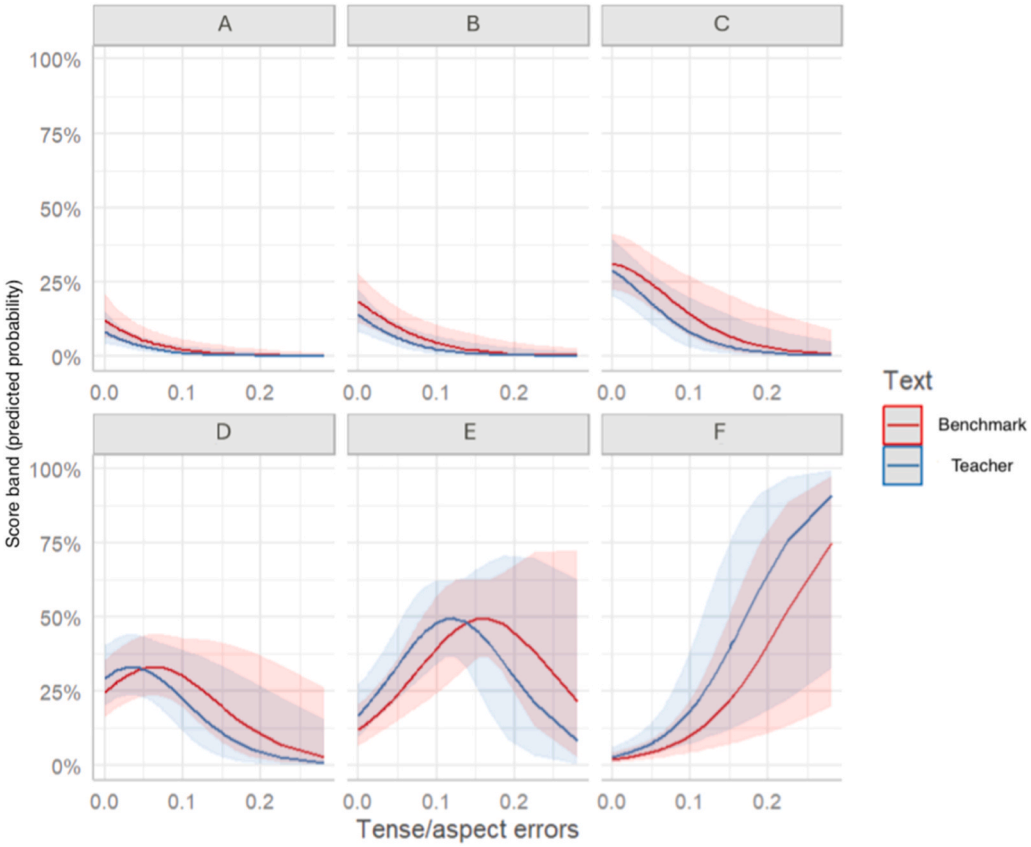


Fig. 6. The effect of misuse of tense/aspect on grading.

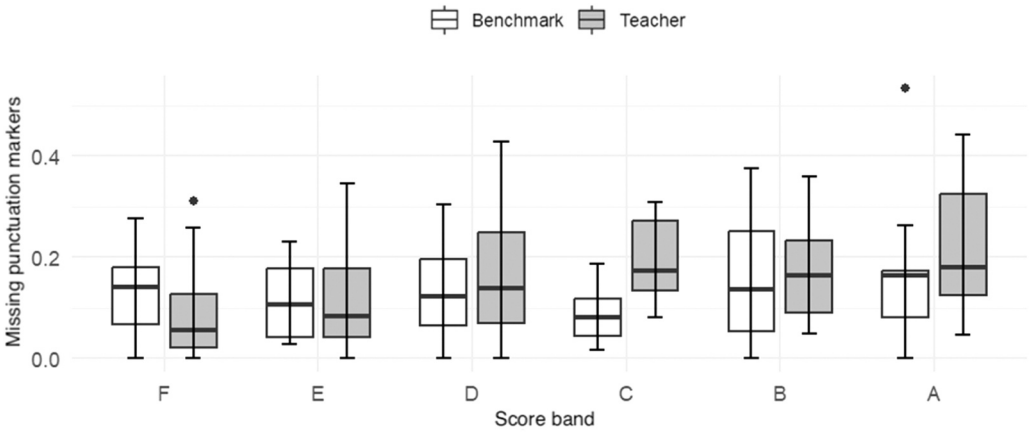


Fig. 7. The relation between missing punctuation markers and score bands.

The study adds to our cumulative knowledge by showing that accuracy has a significant influence on scoring of written texts during high-stakes exams even when it is not an explicit component of the assessment instructions. The role of accuracy in writing assessment has previously been somewhat disregarded, but the findings show that errors possibly have a larger impact on score bands than constructs perceived to be more important like lexical complexity. This may be because accuracy is easier to rate instinctively while lexical complexity requires more training. The results have direct implications for teachers and curriculum/test developers as the former should consider explicitly including accuracy in their teaching practices, while the latter might need to consider how course plans handle accuracy in relation to control instruments such as the national test of English.

This study has several limitations. First of all, for practical reasons, the material was tagged by one person, which is not ideal as

error tagging is a challenging endeavour, and additional taggers could have increased the accuracy of the process. Furthermore, accuracy is only one dimension out of many that affects scoring, and it is likely that other dimensions, such as fluency or content, also affected scoring in the material. Lastly, the number of the teacher scored texts is somewhat small and a larger sample size may have affected the results by increasing the frequency of different error types.

In the future, given the findings of this study and the limitations, it would be desirable to analyse a larger amount of similar material with a group of error annotators, and to include multiple constructs to truly tease out the impact of different dimensions on scores during high-stakes EFL exams. It would also be important to investigate how different language teaching pedagogies explicitly affect students' ability to produce accurate language and what the consequences are, given that students' results, for example, can affect their access to higher education and/or citizenship.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Christian Holmberg Sjöling: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The numerical dataset and code will be made available through OSF if the manuscript is accepted for publication. Access to the analysed corpus may be granted upon request depending on the inquiry.

References

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: does it transfer to a new task? *The Language Learning Journal, 39* (3), 269–280.

Bergström, D., Norberg, C., & Nordlund, M. (2022). Words are picked up along the way"-Swedish EFL teachers' conceptualizations of vocabulary knowledge and learning. *Language Awareness, 31*(4), 393–409.

Bulté, B., & Housen, A. (2012). *Defining and operationalising L2 Complexity Dimensions of L2 Performance and proficiency: Complexity, Accuracy and Fluency in SLA, 32*, 21.

Choi, I., & Cho, Y. (2018). The impact of spelling errors on trained raters' scoring decisions. *Language Education Assessment, 1*(2), 45–58.

Christensen, R. (2023). *Ordinal—Regression Models for Ordinal Data*. R package version, 2023.12-4.1. ⟨https://cran.r-project.org/package=ordinal⟩.

Corder, S.P. (1967). The significance of learner's errors.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Research Report Series, 2005(1), i*, 77.

Díaz-Negrillo, A., & Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista española Délelőtt lingüística aplicada, 19*, 83–102.

Dobrić, N. (2024). Effects of errors on ratings of writing performances – Evidence from a high-stakes exam. *Assessing Writing, 59*, Article 100806. https://doi.org/10.1016/j.asw.2023.100806

Educational Testing Service (2023Writing Scoring Guide. TOEFL iBT® Writing Scoring Guide Flyer(ets.org).

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*(4), 474–509.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Erickson, G., Borger, L., & Olsson, E. (2022). National assessment of foreign languages in Sweden: A multifaceted and collaborative venture. *Language Testing, 39*(3), 474–493. https://doi.org/10.1177/02655322221075067

Erickson, G., & Tholin, J. (2022). Overall, a good test, but…—Swedish lower secondary teachers' perceptions and use of national test results of English. *Languages, 7* (1), 64. https://doi.org/10.3390/languages7010064

Ferris, D. R., & Hedgcock, J. S. (2023). Teaching L2 composition: Purpose. *process, and practice* (4th ed.). Routledge.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics, 36*, 98–116.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heidelberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., & R-Core. (2012). *Package 'car' version, 16* p. 333). *Vienna*: R Foundation for Statistical Computing. ⟨https://cran.uni-muenster.de/web/packages/car/car.pdf⟩.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9–27.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM* (2nd ed.). Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Swallow, H., & Thewissen, J. (2022). The Louvain Error tagging Manual. Version 2.0. CECL Papers 4. *Louvain-Louisiana-Neuve: Centre for English Corpus Linguistics/Université Catholique Délelőtt Louvain (38 Pages) Available here*.

Granger, S., Thewissen, J., & Swallow, H. (2023). The UCLouvain Error Editor User guide - Version 2.0. CECL Papers 6. *Louvain-Louisiana-Neuve: Centre for English Corpus Linguistics/Université Catholique Délelőtt Louvain (33 Pages) Available here*.

Holmberg Sjöling, C. (2025). The effect of lexical complexity on grading of Swedish EFL learners' texts during high-stakes exams. *International Journal of Learner Corpus Research, 11*(2), 245–275.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and Proficiency* (pp. 1–20). Complexity, Accuracy and Fluency in SLA (John Benjamins).
Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. https://doi.org/10.1093/applin/amp048
Johansson, C., & Geisler, C. (2009). The Uppsala Learner English Corpus: A new corpus of Swedish high school students' writing. *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics* (pp. 181–190). *Uppsala*: Acta Universitatis Upsaliensis.
Kleiber, C., Zeileis, A., & Zeileis, M. A. (2020). R package 'aer' version, 1.2-14. *org)*. AER: Applied Econometrics with R (r-project).
Köhlmyr, P. (2003). To err is human. An investigation of grammatical errors in Swedish 16-year-old learners' written production in English. Acta Universatatis Gothoburgensis, Göteborg, Sweden.
Laerd Statistics. (n.d.). Ordinal Regression using SPSS Statistics. How to perform an Ordinal Regression in SPSS | Laerd Statistics.
Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579–589.
Mangiafico, S. (2024). rcompanion: Functions to Support Extension Education Program Evaluation. Rutgers Cooperative Extension, New Brunswick, New Jersey. version 2.4.36. ⟨https://CRAN.R-project.org/package=rcompanion/⟩.
Milton, J., & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. *Entering text*, 127–143.
Neff, J., & Bunce, C. (2006). Pragmatic word order errors and discourse-grammar interdependence. In *IV Conferencia de Lingüística Contrastiva. Santiago de Compostela: Universidad de Santiago.*
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.
Pallotti, G. (2010). Doing interlanguage analysis in school contexts. *EUROSLA Monographs Series, 1*.
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145.
Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161–179.
Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101–143.
R Development Core Team. (2023). *R: A language and environment for statistical computing. [Computer software]*. Vienna, Austria: R Foundation for Statistical Computing. ⟨https://www.R-project.org/⟩.
Schlegel, B.E., Steenbergen, M., Schlegel, M.B., & Imports, M.A. S.S. (2018). R Package brant, version 0.3-0. brant: Test for Parallel Regression Assumption (r-project.org).
Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and Proficiency*. Complexity, Accuracy and Fluency in SLA, 199). John Benjamins.
Skolverket. (2022). *Ämnesplan i Engelska*. Skolverket.
Thewissen, J. (2015). Accuracy across proficiency levels: A learner corpus approach. *Presses Universitaires Délelőtt Louvain, 2*.
Thewissen, J. (2021). Accuracy. In N.T-V. & M. Paquot (Eds.). *The Routledge Handbook of Second Language Acquisition and Corpora*, 305–317.
University of Gothenburg (2023Exempel på texter bedömda som E, Engelska 5. Exempel på uppgiftstyper för Engelska 5 | Projektet Nationella prov i främmande språk - Nafs, Göteborgs universitet (gu.se).
University of Gothenburg (2023Översikt – Skriftlig produktion och interaktion, Engelska 5. Exempel på uppgiftstyper för Engelska 5 | Projektet Nationella prov i främmande språk - Nafs, Göteborgs universitet (gu.se).
Venables, W.N., & $1 $2 (2002). Modern Applied Statistics with S. (4th ed.) Springer.
Wolfe-Quintero, K. E., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. *Technical report. University of Hawai'i, Second Language Teaching and Curriculum Centre, Honolulu*, (17).

Christian Holmberg Sjöling is a 4th year PhD student in English didactics at Luleå University of Technology.