frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 01 August 2025
DOI 10.3389/feduc.2025.1605658

OPEN ACCESS

EDITED BY
Elena Jiménez-Pérez,
University of Malaga, Spain

REVIEWED BY
Antonio Martín-Ezpeleta,
University of Valencia, Spain
Jiu Li,
Shanxi Normal University, China

*CORRESPONDENCE
Flavio Lötscher
✉ flavio.loetscher@phzh.ch

RECEIVED 03 April 2025
ACCEPTED 27 June 2025
PUBLISHED 01 August 2025

CITATION
Lötscher F, Trüb R, Lohmann J, Möller J, Jansen T and Keller SD (2025) Development of grammatical and lexical skills in argumentative EFL writing at upper secondary level in Germany and Switzerland. *Front. Educ.* 10:1605658. doi: 10.3389/feduc.2025.1605658

Development of grammatical and lexical skills in argumentative EFL writing at upper secondary level in Germany and Switzerland

Flavio Lötscher[1]*, Ruth Trüb[2], Julian Lohmann[3], Jens Möller[4], Thorben Jansen[3] and Stefan Daniel Keller[1]

[1]Zurich University of Teacher Education, Zürich, Switzerland, [2]University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland, [3]Leibniz Institute for Science and Mathematics Education, Kiel, Germany, [4]Kiel University, Institute for Psychology of Learning and Instruction, Kiel, Germany

The ability to write argumentative essays is an important requirement in EFL curricula in Germany and Switzerland, with grammar and lexis serving as indispensable elements of this task. The literature shows that acquiring advanced grammatical and lexical skills is challenging for students, especially genre-specific lexical abilities. In this longitudinal study, we investigate how grammatical and lexical skills develop in two educational systems among learners at upper secondary schools (operationalized as number of grammatical and lexical errors). Based on texts from n = 470 learners at two time points at the beginning and end of Year 11, it shows that learners in both countries made more lexical than grammatical errors (partial η^2 = .17). There were no differences between the countries. Longitudinal analyses revealed moderate positive developments in both grammatical and lexical skills over the course of one school year (partial η^2 = .08). The study confirms the importance of vocabulary for advanced L2 writers, which seems to pose a larger challenge than grammar and warrants special attention in EFL writing at upper secondary school. Implications for teacher education and classroom practice such as the emulation of model texts are discussed, with a focus on lexical chunks typical of argumentative writing.

KEYWORDS

English as a foreign language, longitudinal study, upper secondary school, argumentative essay, grammatical skills, lexical skills

1 Introduction

English serves as a global lingua franca across various fields and activities, including science, academia, economics, tourism, and social media (Keller et al., 2020; MacKenzie, 2013). Consequently, curricula in Germany and Switzerland require students to reach a minimum proficiency level of B2 (independent user) according to the Common European Framework of Reference for Languages (CEFR) in both speaking and writing by the end of upper secondary education, indicating an ability to communicate effectively on a range of topics (Bildungsdirektion des Kantons Zürich, 2017; Council of Europe, 2001; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014). EFL writing confronts learners with unique challenges, namely knowing about different text genres, the ability to structure content logically, and sufficient proficiency in writing mechanics (De Smedt et al., 2022; Keller et al., 2024). Within the different genres mentioned in the curricula, argumentative essays hold a special place in

EFL education at this level. Mastery of this genre is often seen as a prerequisite for tertiary education and features prominently in standardized language exams such as TOEFL or CPE (De Smedt et al., 2022; Fleckenstein et al., 2019).

Looking at empirical research examining students' language skills, we find that most extant studies at upper secondary level focus on receptive rather than productive competencies, and that longitudinal studies are particularly rare (Köller et al., 2019; Landrieu et al., 2022). This is largely due to the difficulty and expenses of rater-based scoring (Keller et al., 2020). Moreover, previous empirical investigations in Germany and Switzerland predominantly used holistic rubrics as part of the evaluation process (Keller et al., 2020), which do not allow a differentiated look at individual aspects of the writing construct such as grammatical or lexical skills. A significant research gap is thus to understand the longitudinal development of key text features in EFL writing. For effective teaching in Germany and Switzerland, it is crucial to understand both the curricular requirements and the learners' capabilities.

Keller et al. (2024) looked at *language quality*, *content* and *structure* as three main components of argumentative essays. They found that *language quality*, operationalized as a combination of grammar, vocabulary, and spelling, was the most difficult for learners compared to *content* and *structure*, and that it developed most slowly over the course of one school year. They also found that Swiss students outperformed their German peers, despite Germany's centralized approach focusing on argumentative structure and linguistic accuracy, while Switzerland employs a decentralized, literature-focused framework (Keller et al., 2024; see Section 2.4). Communicative approaches as well as "competence based" curricula (Richards and Rodgers, 2014) shifted attention in writing instruction away from grammar- and vocabulary-learning in isolation. However, linguistic quality remains a critical factor in EFL writing at this level, as grammar and vocabulary errors often obstruct comprehension and distract readers from the content (Biesenbach-Lucas, 2007). Furthermore, at upper secondary level, prioritizing linguistic accuracy seems well justified, as students must meet the advanced language demands of tertiary education and standardized exams, which require a strong command of grammar and lexis to succeed without relying on AI or writing assistant tools, as emphasized in the curricula. The objective of this essay is therefore to take a closer look at the aspect of language quality. We focus on developments at the level of grammar and lexis as these are key features of writing quality and prerequisites for higher hierarchy aspects such as argumentation, content and organization.

For this purpose, we have structured this paper into four parts, as follows: In the background section, we summarize relevant studies on grammar and lexis in advanced EFL argumentative writing, followed by research on their influence on reader perception, which plays a key role in the assessment and evaluation of writing quality. Next, we provide an overview of the educational contexts in Germany and Switzerland, in which this study was conducted. In the methods section, we outline the development of the analytic rating system, rater training and statistical analyses conducted in this study. In the results section, we present a detailed account of two aspects of text quality (grammar and lexis) by contrasting the learners' competencies and their development over one year in both countries. In the final section, we discuss the relevance of the results for teaching argumentative EFL writing, along with some limitations and perspectives for further research.

2 Background

2.1 Key components of argumentative essays

From a learner's perspective, receptive and productive foreign language skills are closely connected, as they are usually required together in real-life communicative contexts. However, from a research perspective, it is often necessary to isolate a skill to ensure analytical clarity and methodological control. This tension is also reflected in the evolution of foreign language pedagogy. According to Richards and Rodgers (2014), the introduction of the communicative approach in the early 1980s marked a far-reaching paradigm shift and competence orientation has since become the basis for EFL teaching in Western European countries such as Germany and Switzerland (Bildungsdirektion des Kantons Zürich, 2017; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014). The curricula focus on what learners can do with the language and put more emphasis on content and language in use, rather than targeting a native speaker-like accuracy. While this approach has triggered many positive developments, it also created some backlash, with teachers for example ignoring emails from students that had not been checked for language errors (Biesenbach-Lucas, 2007). This inconsistency becomes even more challenging in argumentative writing in the EFL context, where the presentation of a convincing argument depends upon learners having a solid grasp of aspects such as grammar and lexis. In accordance with the Common European Framework of Reference for Languages (CEFR), the more complex the genre and the more advanced learners' EFL competencies, the more need there is for highly developed linguistic competencies at the level of language mechanics, lexis and syntax (Cumming et al., 2002; Thornbury, 1999). The importance of language quality in argumentative writing has been confirmed in several studies (Cumming et al., 2002; Hyland, 2003; Rezaei and Lovorn, 2010). Hyland (2003), for example, observed that despite teachers' beliefs and teaching approaches, *language quality* remained the central focus in their assessments. Similarly, Rezaei and Lovorn (2010) noticed that the overall essay evaluation was more strongly influenced by grammatical, syntactical and spelling issues than by content, even when instructing raters to disregard mechanical details.

Analytic rubrics typically feature grammatical and lexical skills, together with language mechanics, as central elements of linguistic writing quality (Cushing, 2019). The distinction between grammar and lexis was also part of Jacobs et al.'s ESL Composition Profile (1981), even though they opted for a different label when naming the traditional grammatical attributes *language use*. Since then, similar categories have been used in follow-up studies (cf. Kim, 2011) and Henry (2000) mentioned that grammar and vocabulary were two indispensable skill sets of essay writing. The complementary nature of the two has also been incorporated into the curricula cited above (Bildungsdirektion des Kantons Zürich, 2017; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014). The curriculum in

Schleswig-Holstein, for example, states that "with regard to vocabulary and grammar, students can use their lexical resources in a context-oriented manner" and "apply grammatical structures to support their speaking and writing intentions" (Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014, p. 23). The development of grammar and lexis thus remains an important field of study, in particular in order to understand how to support learners.

2.1.1 Studies on grammar in EFL writing

An overview of relevant studies shows that there are widely varying operationalizations of the grammar construct. Typical operationalizations include verb usage, tenses, subject-verb agreement and word order (Darus and Subramaniam, 2009; Keller et al., 2024; Lahuerta, 2018; Zhan, 2015). Other grammatical structures examined include plural formations (Darus and Subramaniam, 2009) and the correct usage of pronouns (Keller et al., 2024). When it comes to the categorization of prepositions, the issue becomes inconsistent. Even though they are typically regarded as lexical items (Keller et al., 2024; Nuruzzaman et al., 2018), Wolf et al. (2018) classified prepositional errors as grammatical errors. In accordance with Lewis' lexical approach (1993), this variation seems natural as grammar is a multi-faceted construct containing different syntactical phenomena and to some degree overlaps with lexical phenomena (e.g., multi-word items). Lewis (1993) also argues that grammar and lexis are distinct yet interconnected aspects of language teaching and suggests that EFL instruction should prioritize lexis over grammar, integrating grammar within a lexical framework for effective communication.

It has been shown that different grammar aspects hold different challenges for learners. Nuruzzaman et al. (2018) identified tense selection as the most error-prone issue for Saudi students, followed by subject-verb agreement. Darus and Subramaniam (2009) similarly observed that four of the six most frequent problems among Malaysian participants involved grammar, with errors in tense choice, subject-verb agreement, plural forms, and word order being most common. Zhan's (2015) study of Chinese EFL writers also highlighted errors in tense and verb forms as the top issues in topic-based writings. Among Spanish learners, the majority of errors involved verb tense issues, particularly with modal auxiliaries, and grammatical morphemes such as the omission of suffixes, as noted by Lahuerta (2018).

2.1.2 Studies on lexis in EFL writing

The value of vocabulary for EFL writing – and foreign language learning in general – has long been acknowledged (Barkaoui, 2010). Similar to the operationalization of grammar, past studies used different conceptualizations of lexical features. Read (2000) distinguished between *density* (i.e., proportion of content words compared to function words), *diversity* (i.e., use of different types of words), *sophistication* (i.e., use of advanced words), and *accuracy* (i.e., number of errors). Importantly, the concept of *lexical accuracy* too has been classified differently. A well-known classification was presented by James (1998) when he distinguished between *mis-selection* (wrong word choice), *mis-formation* (words that are non-existent in the L2 but exist in the L1) and *distortion* (words that are non-existent in both the L2 and the L1). As mentioned above, Keller et al. (2024) and Nuruzzaman et al. (2018) both considered prepositional and article errors as lexical errors, while orthographic shortcomings were dealt with in a separate category. Llach (2011), on the other hand, also included misspellings in the lexical category.

Having addressed the varying definitions of lexis and lexical accuracy, it is important to assess their impact on EFL writing quality. Apart from content, lexical quality was recognized as the most central attribute of high-level essays (Schoonen et al., 2009). Even though vocabulary and content mostly represent separate components in assessment rubrics (cf. Cushing, 2019; Jacobs, 1981), the former, according to Harmon et al. (2005), indirectly reveals the writer's knowledge as many topics use specialized terminology. Various studies show that to predict the overall text quality, the most promising strategy is to determine the level of word choice (Barkaoui, 2010; Ferris, 1994; Song and Caruso, 1996). Swan (1988) considered vocabulary one of the most difficult aspects for EFL students, with prepositional phrases being of particular importance. Looking at accuracy, studies on lexical errors were traditionally less common than studies on grammatical errors as they were considered more difficult to systematize, classify, generalize, and remedy (Llach, 2017). Sermsook et al. (2017) aimed to analyze language errors in the writing of English major students at a Thai university, and their findings showed that, after punctuation, article usage was the most common area of error. The importance of vocabulary has also been recognized by the learners. In a study of $n = 128$ undergraduates, Leki and Carson (1994) found that students regarded vocabulary instruction as the area where they required the most support. Similarly, Božić Lenard et al. (2018) showed that a majority of EFL students considered vocabulary assignments more appealing, more useful, and more worthy of their study time, while their grammatical skills were significantly more advanced.

2.2 Longitudinal studies of grammatical and lexical development

Beyond the importance of grammar and lexis for writing quality, a central follow-up question concerns their development over time. As Barkaoui and Hadidi (2020) point out in their overview, the vast majority of studies on L2 writing skills were cross-sectional, and longitudinal designs still represent a marginal portion of the scientific discussion. Since the ability to produce high-level texts includes various traits that evolve at unequal paces (Weigle, 2002), more longitudinal research is needed (Kim, 2021). Yet, another challenge is the fact that most longitudinal examinations were case studies (Macqueen, 2012). When investigating a larger number of participants, the focus was typically placed on *lexical sophistication* or *syntactical complexity* (Barkaoui and Hadidi, 2020; Kim, 2021). Of all the longitudinal studies summarized prior to their own contribution, Barkaoui and Hadidi (2020) only identified a few studies that, among other features, explored changes in accuracy. Knoch et al. (2014) followed 101 students from various Asian countries over the course of one school year and found that only their fluency had improved. In a study of $n = 58$ participants, Polio and Shea (2014) found that a 15-week intervention improved the holistic scores of *language use* and *vocabulary*, but did not enhance scores in *accuracy*. Finally, Barkaoui and Hadidi's 85 Chinese participants (2020) made significantly fewer errors after their guidance period of 6–9 months, but raters had been instructed to consider all types of errors jointly (lexical, morphosyntactic, semantic and mechanical) before making an overall judgment as to the overall impact of these errors (pp. 40–41, 169). These studies have shown that both grammar and lexis are integral

components of the developmental process but were often not analyzed as separate elements. Two exceptions were Yoon's study (2018), which involved 51 participants from diverse L1 backgrounds, and the research by Storch and Tapper (2009), which primarily focused on Chinese postgraduate students. Yoon's study (2018) found a significant increase in vocabulary scores but not in grammar scores when assessing complexity and errors together. Storch and Tapper's research (2009), on the other hand, observed a marked improvement in academic vocabulary use by the end of an English for Academic Purposes (EAP) course, as measured by Coxhead's Academic Word List (Coxhead, 2000). Moreover, participants demonstrated improved grammatical accuracy over time. The categories of errors included syntax (e.g., word order errors, missing elements), morphology (e.g., verb tense, subject-verb agreement), grammar (e.g., articles, prepositions), and lexis (e.g., word choice). In a more recent intervention study, Wu et al. (2023) demonstrated that using model texts is a promising writing instruction approach in the classroom, as it significantly improved the writing skills of their Chinese EFL learners ($n = 60$), leading to sustained enhancements in both grammatical and lexical quality over a one-week period.

2.3 Impact of grammar and lexis on the assessment of argumentative writing

In the preceding sections we have outlined the importance of grammar and lexis for the EFL writing construct and their development over time. This section examines the impact of those factors on how raters perceive a text, with particular emphasis on how language quality influences comprehension and affects judgments regarding argumentative strength and overall quality. What raters think are key aspects and the qualities they mostly consider when assessing students' essays often differ fundamentally (Breland and Jones, 1984). For example, McNamara (1996) found that even though their study participants had estimated the influence of grammar to be low, grammatically correct essays constantly received higher scores on holistic scales. Vögelin et al. (2019) discovered that the manipulation of lexical quality not only led to changes in the rating of a text's vocabulary, but simultaneously influenced the evaluation of grammar and frame of essay (i.e., presence of introduction and conclusion), which is known as a halo effect. Fritz and Ruegg (2013) discovered that when assessing vocabulary, it was the number of errors (not sophistication or diversity) that was most responsible for the scores as under time pressure, raters would concentrate on the most salient feature.

Some studies identified raters' L1, experience, and the form of feedback as moderating variables for assessment outcomes. Lee (2016) showed that for experienced raters, content was the most crucial element whereas for novice raters, vocabulary was more essential. Similarly, Higgs and Clifford (1982) found that students' lexical skills were more important for less experienced raters while the grammatical competence was what experts valued the most. Song and Caruso (1996) discovered that native speaking raters put more emphasis on content while non-native speakers paid more attention to grammatical accuracy and were generally more critical of grammatical errors. Another insightful detail regarding grammar was discovered by Weigle et al. (2003), who found that it was the overall quality of the text that changed its role, with content being most important in stronger texts and grammar gaining in importance in weaker texts. Grammar and lexis, therefore, generally play a major role in shaping how raters perceive and assess texts. This strong influence reinforces the notion that helping students to express meaning and formulate arguments with appropriate grammar and lexis remains a key element in teaching EFL writing at upper secondary school. For teacher instruction and classroom practices in Germany and Switzerland, it is essential to know the requirements of the curricula on the one hand and what learners are capable of achieving on the other hand.

2.4 Context of EFL writing in Germany and Switzerland

Regarding the English competencies which students are expected to achieve by the end of upper secondary school, the curricula in Germany and Switzerland both adhere to the CEFR standards (Council of Europe, 2001) and stipulate level B2 (Bildungsdirektion des Kantons Zürich, 2017; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014). At this level, argumentative writing demands students being able to deliver clear, detailed accounts on numerous topics, to explain their point of view on various issues, and to assess the advantages and disadvantages of different options related to a specific prompt (Council of Europe, 2001). When grammar and lexis is concerned, the CEFR standards and the syllabi in both countries specify that learners should display good grammatical control with minor errors, a broad vocabulary covering general and field-specific topics, flexibility in expression, appropriate use of collocations, and high lexical accuracy, with occasional challenges in complex structures and specialized terms outside their field (Bildungsdirektion des Kantons Zürich, 2017; Council of Europe, 2001; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014). Even though the requirements at upper secondary school in Germany and Switzerland are almost identical, their evaluation processes are fairly different. Germany provides a set of national standards that include a variety of assessment tasks employed nationwide in the final examinations. These guidelines suggest a focus on argumentative writing and there are national rubrics intended for the assessment of *structure* and *language quality*. Consequently, a particular emphasis is given to text cohesion, grammatical and lexical features (Institute for Educational Quality Improvement [IQB], 2021). In Switzerland on the other hand, the focus of teaching English at this level usually lies on interpreting literary works, which play a central role in the final exams (Keller, 2013). There are no binding forms of assessment for argumentative EFL writing at a national level. Instead, it is the duty of the cantons and schools to regulate how much weight is given to persuasive essays and how they are tested or evaluated. Whether this divergence would lead to varying student competencies was first analyzed by Keller et al. (2020) because prior to that, no publication had verified that setting B2 as a goal was realistic or appropriate. Keller et al. (2020) showed that one year before graduation, more than 70% of the students achieved the required minimum in both countries. When comparing the two educational systems, it was discovered that the Swiss students generally outperformed their German peers, regardless of the rating method (i.e., holistic or analytic) (Keller et al., 2020, 2024). A closer look at the analytic ratings revealed that the most significant differences could be attributed to the internal

structure of the essays, where the German learners initially exhibited some systematic weaknesses but later on progressed more quickly (Keller et al., 2024). No such differential developments were observed for *content* or *language proficiency* (Keller et al., 2024). While it was demonstrated that Swiss students outperform their German peers both holistically and in terms of structure, it remains unknown whether new differences emerge when dismantling the weakest and slowest-developing component – language quality – and thoroughly examining grammar and lexis.

2.5 Research questions

As outlined in the previous sections, language quality was the most challenging and slowest-progressing feature for German and Swiss students, while remaining a top priority for raters. This makes it essential to disentangle it at the level of grammar and lexis – its two core components in the respective curricula. Prior research has identified significant differences between Germany and Switzerland in holistic ratings and structural aspects (Keller et al., 2020, 2024; see Section 2.4), but it remains unknown whether the different educational systems also contribute to variations in grammatical and lexical competencies. Finally, longitudinal writing research is scant and often limited by small sample sizes. In the light of these gaps, we formulated the following research questions for our study:

Differences between aspects of language quality

1. Are there differences between the grammatical and the lexical skills of German and Swiss learners when writing English argumentative essays at upper secondary school?

Differences between educational systems

2. Are there differences between German and Swiss learners with respect to the grammatical and lexical skills when writing English argumentative essays at upper secondary level?

Longitudinal development

3. Do German and Swiss learners at upper secondary school improve their grammatical and lexical skills within the course of one school year when writing English argumentative essays?

3 Research methods

3.1 Samples and procedures

All data was collected by Keller et al. (2020) and carried out as a repeated measurement design in upper secondary schools in Germany and Switzerland with an interval of approximately 9 months between time point 1 (=T1; August/September 2016) and time point 2 (=T2; May/June 2017). To operationalize EFL-argumentative writing skills, the authors used two prompts from the TOEFL-iBT® writing section (Educational Testing Service, 2009). They were chosen because of the alignment between the curricula in Germany and Switzerland and the holistic evaluation grid of the TOEFL independent task, which requires learners to state, explain, and support their opinion on a controversial topic without prior input as in the integrated task (cf. Fleckenstein et al., 2019). The following two prompts were used: "A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught" (Teacher, or TE); and "Television advertising directed toward young children (aged two to five) should not be allowed" (Advertising, or AD). Study participants were instructed to explain reasons for their points of view in an argumentative essay of approximately 300 words (cf. Keller et al., 2020). The writing time was 30 min without any preparation or aid. The authors provided laptops in a rotational matrix design where the order in which the learners worked on the prompts at T1 and T2 was randomly varied to exclude possible sequence effects (cf. Rupp et al., 2019), while task comprehension was ensured through direct supervision and monitoring during test administration. In Germany, data was gathered in the federal state of Schleswig-Holstein, where 42 schools were randomly selected from a pool of 84 schools and students were asked for their voluntary participation. In the end, $n = 965$ students from various subject tracks (e.g., language, science, civics) participated in the study (58.60% female; age $M_{T1} = 16.91$, $SD_{T1} = 0.56$; $M_{T2} = 17.61$, $SD_{T2} = 0.56$). In Switzerland, 20 schools from the following seven cantons agreed to participate: Aargau, Basel Stadt, Basel Land, Luzern, St. Gallen, Schwyz and Zürich (all of these cantons belong to the German-speaking part of Switzerland, where Swiss German is the local language). Data was collected from 91 classes with various specialist subjects (e.g., Latin, music, visual arts), resulting in a sample size of $n = 1882$ Swiss students (58.00% female; age $M_{T1} = 17.56$, $SD_{T1} = 0.91$; $M_{T2} = 18.27$, $SD_{T2} = 0.91$). Although the subject tracks in Germany and the special subjects in Switzerland are not perfectly equivalent in terms of curricular scope, they represent the most appropriate basis for ensuring comparability between the two countries, especially given that both samples included learners from a range of educational profiles. For the analytic analyses of this study, students who had not partaken in both measurement time points were excluded, resulting in a final sample of $n = 159$ students from Germany (59.20% female; age $M_{T1} = 16.68$; $M_{T2} = 17.42$, $SD = 0.58$) and $n = 311$ students for Switzerland (63.70% female; age $M_{T1} = 17.38$, $M_{T2} = 18.13$, $SD = 0.87$).

3.2 Analytic assessment of learner texts

Since previous analyses of students' grammatical and lexical skills paid less attention to *accuracy* as summarized in Section 2.2, we operationalized the respective skills as the number of errors,[1] which is more narrow than Read's constructs of *diversity*, *sophistication*, and *density* (2000). However, we believe that this offers two major advantages: One is its transferability into school practice as mentioned in Section 2.1 and the other one its economy and reliability of scoring. We purposely separated two aspects of writing quality (grammar and

1 Brown (1994, p. 205) defines a mistake as a slip or failure to apply a known rule, and an error as a systematic deviation from adult native grammar, reflecting a gap in knowledge. In this study, we did not distinguish between the two terms, as the causes of the learners' shortcomings were unknown.

vocabulary) that are often assumed to be related (Lewis, 1993). Our primary reasons for separating the two constructs in this analysis are as follows: not only do the curricula in Germany and Switzerland explicitly list them as separate proficiency requirements (Bildungsdirektion des Kantons Zürich, 2017; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014), but practical experience shows that they are commonly separated in many upper secondary schools in both countries. The research team developed a rater manual to specify the types of errors to be considered when rating the texts (see Table 1). For grammar, we decided to consider mainly verb issues (tenses, negations and modals), but also errors in apostrophizing, relative pronoun use, noun countability, and comparative/superlative adjective forms. In terms of lexis, the use of German words, superfluous/missing words and incorrect prepositions/articles were categorized as lexical errors. We decided to classify prepositional errors as lexical errors, following the English essay grading guidelines of the Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (2021). We applied this approach because in school contexts prepositions are typically treated as vocabulary items due to their frequent collocations with verbs, nouns, and adjectives.

The counting process was shortened if there were more than seven errors, and the respective texts were allocated to the lowest skill level. We excluded spelling errors from the analysis as they did not fall within the defined categories and were deemed insignificant to our research questions.

TABLE 1 Grammatical and lexical aspects included in textual analysis (with examples).

Grammatical aspects	
Descriptor	**Examples of errors**
Verb tense	As I **have seen** last year
Negation without auxiliary	Someone who **has not** the knowledge
Conjugation	He **do not**
Wrong grammatical suffix	They should not allow**ed** to do that
Apostrophizing	If **its** good or bad
Relative pronoun	The person **which**
Noun countability	Their **feedbacks are** important
Comparatives and superlatives	This is **difficulter**

Lexical aspects	
Descriptor	**Examples of errors**
Wrong choice of word	The ability to relate to them is a good **point** for the students
Superfluous word	This feeling gives the teacher **will** the ability to relate
Missing word	This can understood by the _ of students
German word	If the teacher has a lot of **pädagogisches** knowledge
Wrong preposition	Most children must go **in** school
Wrong article	**The** society needs to do something against it

Examples for each grammatical and lexical error category are printed in bold.

3.3 Rating procedures

To receive detailed and reliable scores for all essays regarding the two constructs under investigation (grammatical and lexical skills), we collaborated with the Data Processing and Research Center (DPC) in Hamburg, which is a member of the International Association for the Evaluation of Educational Achievement (IEA). Under the leadership of one co-author, all rating processes were planned and carried out by DPC in cooperation with the research team. We hired seven experts from a pool of employees who had previously completed similar evaluations of authentic student texts. The raters possessed a high level of English proficiency and were aware of the agency's assessment procedures (cf. Keller et al., 2024). Since our goal was to mitigate differences in rater severity and consistency as achieved in earlier studies (Cushing, 2019), the preparations for the rating processes involved a series of training sessions in which the raters were familiarized with the assessment framework. To create a shared understanding of the text qualities, prototypical excerpts for the different levels of proficiency were identified and analyzed. Between the training sessions, raters were instructed to use the analytic criteria and assess ten texts from a trial sample. During the meetings, raters discussed their evaluations in detail, leading to adjustments in the descriptions of the various sub-elements and the removal of inconsistent categories from the rater manual. Due to the restrictions of the COVID-19 pandemic, all exchanges happened online and were led by the research team together with an experienced head rater from DPC.

3.4 Interrater reliability

After the comprehensive training sessions, each text was independently assessed by two experts. As described in Section 3.3, the initial pool consisted of seven raters. The first round of rater training revealed that the interrater reliability was altogether sufficient, but not consistently above the threshold of quadratic weighted Kappa > .60. When reviewing all scores individually, we detected consistently varying results from one rater, who was excluded from further analyses. The subsequent ratings were carried out in pairs of two from the pool of the remaining six raters. This step improved the interrater reliability and lifted the values to a more acceptable level, with .70 for grammatical errors and .60 for lexical errors. The reliabilities for both features were consistent across both writing prompts. The confusion matrices for all rater pairs across both prompts can be found in Appendix 1. Two separate many-facet Rasch analyses (Eckes, 2015) for grammar and lexis were run to adjust the scores for rater severity and task difficulty. By anchoring the facet measurement time point at zero, we could extract corrected mean values for every student at both measurement time points. According to Linacre (2018), the data fit the model when about 5% of the absolute standardized residuals are equal to or greater than 2, and about

1% is equal to or greater than 3 (p. 167). The grammar analysis showed that 3.30% of the responses had an absolute standardized residual equal to or greater than 2 and 0.27% an absolute standardized residual equal to or greater than 3. The lexis analysis showed that 3.78% of the responses had an absolute standardized residual equal to or greater than 2 and 0.64% an absolute standardized residual equal to or greater than 3. These figures, therefore, appear to indicate satisfactory model fits. The grammar analysis showed a 0.85-logit spread for rater severity measures, compared to a 6.75-logit spread for examinee proficiency measures, indicating similar rater severity. Infit and outfit statistics were close to 1 (0.90–1.09) for all six raters, indicating an internally consistent use of the scales. The lexis analysis showed a 1.06-logit spread for rater severity measures, compared to a 4.87-logit spread for examinee proficiency measures, indicating similar rater severity. Infit and outfit statistics were close to 1 (0.84–1.14) for all six raters, indicating a fairly consistent use of the scales. After rounding them to two decimal places, we used the corrected mean values for our analyses.

3.5 Statistical analyses

We ran a three-way mixed ANOVA to examine the effects of feature (grammatical and lexical errors; RQ 1), nationality (Germany and Switzerland; RQ 2) and time (T1 and T2; RQ 3) on the number of errors when writing argumentative essays. Q-Q plots were used to verify normal distribution. If not normally distributed, we applied square root, logarithmic and reciprocal transformations. If the data still deviated from normality after these transformations, we proceeded with the parametric tests nonetheless, given their robustness to deviations from normality in samples larger than 30 (Wilcox, 2021) and their ability to model interactions between factors. To validate our findings, we ran Wilcoxon signed-rank tests for the within-subjects factors *feature* (RQ 1) and *time* (RQ 3) and Mann–Whitney U tests for the between-subjects factor *nationality* (RQ 2).

4 Results

4.1 Descriptive statistics

Table 2 summarizes the performances of all students across both measurement time points. We also include Pearson correlations to show the interrelatedness of error types and to ensure transparency for future research, including potential meta-analyses. In total, there were $n = 940$ essays from 159 German and 311 Swiss students. At T1, the German students had an average score of 3.83 for grammar, which corresponds to approximately 3.8 instances of grammatical errors per text ($SD = 2.19$). Their average score for lexis was 4.46 ($SD = 2.08$), which corresponds to approximately 4.5 lexical errors per text. After one school year, the average scores for the German students were 3.27 for grammar ($SD = 2.20$) and 3.88 for lexis ($SD = 2.07$). Viewed over the entire time period in Germany, the average scores were 3.55 for grammar ($SD = 1.92$) and 4.17 for lexis ($SD = 1.74$).

In Switzerland, the grammar measure at T1 was 3.44 ($SD = 2.03$) and the lexis measure was 4.23 ($SD = 2.01$). After one school year, the Swiss grammar measure was 3.01 ($SD = 1.97$) and the Swiss lexis measure was 3.77 ($SD = 1.93$). Viewed over the entire time period in Switzerland, the average scores were 3.22 for grammar ($SD = 1.65$) and 4.00 for lexis ($SD = 1.63$).

TABLE 2 Descriptive statistics at T1 and T2 in both countries with Pearson correlations.

Germany	*M*	*SD*	1	2	3	4	5	6
1. Grammatical errors T1	3.83	2.19	–					
2. Lexical errors T1	4.46	2.08	.59**	–				
3. Grammatical errors T2	3.27	2.20	.52**	.47**	–			
4. Lexical errors T2	3.88	2.07	.49**	.41**	.47**	–		
5. Grammatical errors T1 + 2	3.55	1.92	–	–	–	–	–	
6. Lexical errors T1 + 2	4.17	1.74	–	–	–	–	.69**	–

Switzerland			7	8	9	10	11	12
7. Grammatical errors T1	3.44	2.03	–					
8. Lexical errors T1	4.23	2.01	.46**	–				
9. Grammatical errors T2	3.01	1.97	.37**	.38**	–			
10. Lexical errors T2	3.77	1.93	.38**	.38**	.37**	–		
11. Grammatical errors T1 + 2	3.22	1.65	–	–	–	–	–	
12. Lexical errors T1 + 2	4.00	1.63	–	–	–	–	.58**	–

$n = 159$ in Germany; $n = 311$ in Switzerland. T1 = time point 1 (August/September 2016); T2 = time point 2 (May/June 2017). Scores are based on the corrected mean number of errors as independently assessed by two trained raters. $^{**}p < .01$.

4.2 Three-way mixed ANOVA

We ran a three-way mixed ANOVA to examine the effects of feature (grammatical and lexical errors; RQ 1), nationality (Germany and Switzerland; RQ 2) and time (T1 and T2; RQ 3) on the number of errors when writing argumentative essays. Inspection of Q-Q plots revealed no normal distribution across all factors and conditions. The distributions remained non-normal after square root, logarithmic and reciprocal transformations. As the non-parametric tests confirm our results, we include them in Appendix 2 and focus on the ANOVA in this section. This allows for an integrated analysis of all factors including their interactions and offers a clearer picture than multiple separate non-parametric tests. Table 3 shows that the main effect of feature (RQ 1) and the main effect of time (RQ 3) were both statistically significant. The essays revealed more lexical than grammatical errors and students reduced the number of errors at T2. The between-subjects factor nationality (RQ 2) was not statistically significant. There were no statistically significant two-way interactions of feature and nationality, feature and time and nationality and time. There was no statistically significant three-way interaction between feature, nationality and time. The interactions between the two within-subjects factors (feature and time) are illustrated in Figure 1 for Germany and Figure 2 for Switzerland. The interactions between the between-subjects factor nationality and the within-subjects factor time are illustrated in Figure 3 for the grammatical errors and Figure 4 for the lexical errors.

Figures 1–4 show that in both countries, students were struggling more with lexis than with grammar and that the improvements in accuracy were moderate. The error count in Switzerland was lower, but this difference was not significant.

5 Discussion

The objective of this study was to examine the quality of students' argumentative essays, with a special emphasis on grammatical and lexical skills as key components of the writing construct. For this purpose, we analyzed argumentative essays at upper secondary level in Germany and Switzerland from a cross-sectional and longitudinal perspective. In particular, we aimed at extending earlier EFL research by providing a deeper understanding of how these two features develop in the specific

TABLE 3 Three-way mixed ANOVA with grammatical and lexical errors as dependent variables.

Measure	$F(1, 468)$	Partial η^2
Feature (RQ 1)	93.02***	.17
Nationality (RQ 2)	2.71	.01
Time (RQ 3)	40.83***	.08
Feature × Nationality	1.21	<.01
Feature × Time	0.03	<.01
Nationality × Time	0.55	<.01
Feature × Nationality × Time	<0.01	<.01

***$p < .001$.

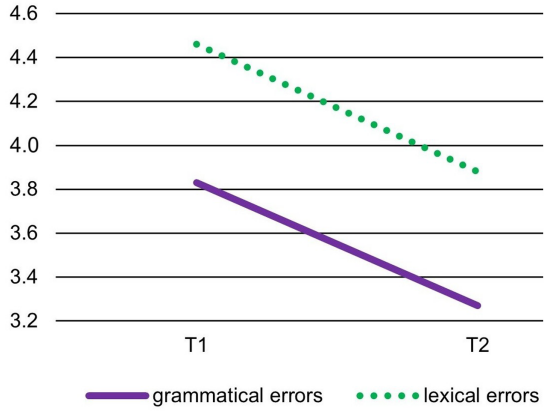

FIGURE 1
Development of writing skills in Germany: grammatical and lexical errors.

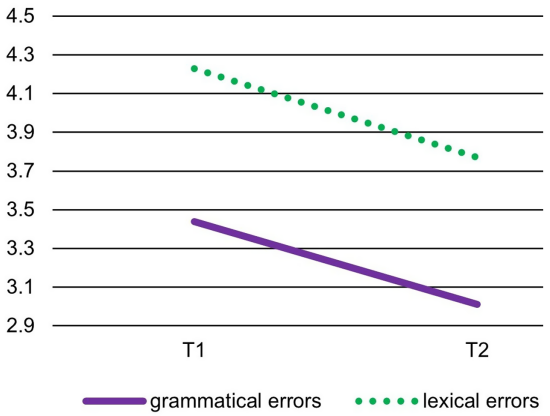

FIGURE 2
Development of writing skills in Switzerland: grammatical and lexical errors.

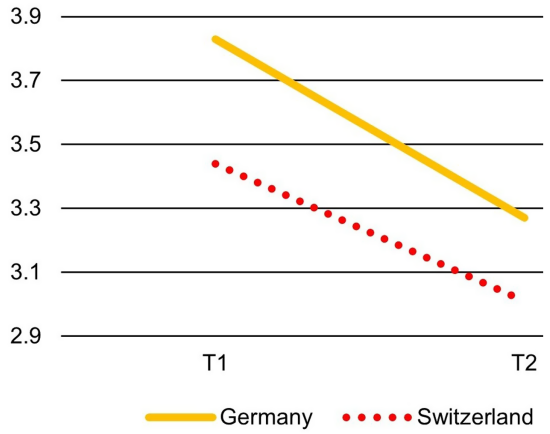

FIGURE 3
Development of grammatical errors in both countries.

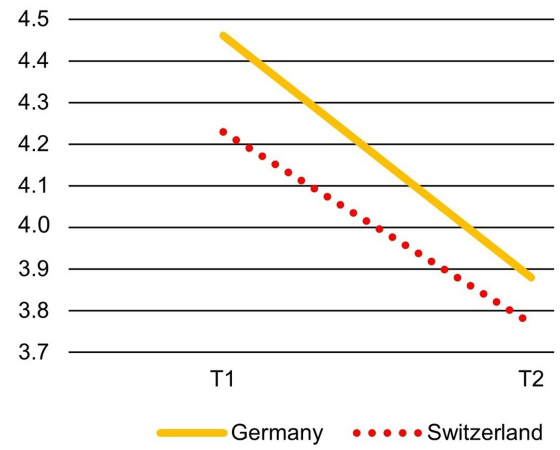

FIGURE 4
Development of lexical errors in both countries.

context of upper secondary education in the two countries. In this section, we discuss the results and outline implications for teaching argumentative writing in the classroom.

5.1 Comparing grammatical and lexical error counts (RQ 1)

Our analysis focused on grammatical and lexical errors as an indicator of text quality and showed a significant difference between the two constructs. We found that lexical errors outnumbered grammatical errors across both countries and measurement points, despite the fact that we classified some theoretically contentious cases (e.g., noun countability) as grammatical rather than lexical. At upper secondary level, students in Germany and Switzerland thus seem to have gained a certain mastery over the English tense system, negations, subject-verb agreement and word order, while struggling more with vocabulary and the correct choice of articles or prepositions for example. Our findings align with earlier international research. Swan (1988) identified vocabulary as one of the most difficult aspects of language learning for EFL students, emphasizing the persistent challenges learners face in acquiring lexical resources. Similarly, Božić Lenard et al. (2018) found that EFL learners' grammar was more advanced than their vocabulary, suggesting a developmental imbalance between the two constructs. Our results thus reinforce the necessity of more specific and targeted instructions in lexical aspects, particularly concerning argumentative word chunks (see Section 5.4).

5.2 Comparing German and Swiss students (RQ 2)

Descriptive results showed that German students displayed slightly lower skill levels as their texts on average contained more grammatical and lexical errors at both measurement points. However, the differences between the two countries were not statistically significant. This does not fully align with previous studies. For instance, Keller et al. (2020) reported that Swiss students performed better on the holistic scale, while Keller et al. (2024) identified an interaction effect at the structural level of the essays, noting that Swiss learners initially outperformed their German peers, with steeper gains observed in Germany. We propose that this difference may only be visible at a macro level when multiple features are combined to assess general competencies such as *language quality* or *structure*. As the focus of this investigation is more narrow, we assume that such differences between the countries, which are generally small, disappear when examining individual features at the micro level. In upper secondary schools, the mastery of grammar and lexis in argumentative essays seems to be equally difficult for students in both educational systems.

5.3 Longitudinal developments (RQ 3)

The results of the current study revealed moderate improvements over the course of one school year. The small but consistent improvements in accuracy show that the educational systems in Germany and Switzerland are generally effective, although at development rates which are rather slow but in line with international studies (Barkaoui and Hadidi, 2020). Some of our findings contradict Yoon's (2018) observations, who reported that his participants had improved the lexical but not the grammatical competencies. However, we suspect that this difference stems from varying operationalizations: Yoon's investigation (2018) did not exclusively focus on error counts but also included *complexity* and *sophistication*. Nevertheless, it seems safe to assume that lexical and grammatical skills are both part of a general language proficiency (Lewis, 1993), which develops more slowly than the ability to build a coherent argument and present it in an introduction, main body, and conclusion (Keller et al., 2024).

5.4 Pedagogical implications

Our results have a range of practical implications, and we would like to present some strategies that can be introduced in the teaching of EFL argumentative writing skills. Despite the interconnection between the lexical approach and grammar as proposed by Lewis (1993), we argue that it can still be beneficial to distinguish between these two aspects when giving feedback and revising argumentative essays. This distinction helps provide learners with instructional support that aligns with their strengths or weaknesses, such as tense selection or the choice of prepositions. Even though essay writing is not solely about avoiding mistakes, they have the potential to highlight gaps in knowledge, misinterpretations, and, if handled purposefully, attention and achievement deficits among students (Hascher, 2005). Our results underscore the importance of text-functional and genre-based writing approaches (Hyland, 1990), which suggest that in order to make fewer mistakes, students need to be familiarized with typical words and word combinations used in argumentative essays. Writing is a teachable skill and increasing the visibility of what is to be learnt must be a vital part of teaching (Hyland, 1990). Model texts can be used for students to analyze and

master the typical lexical structures of a genre. This might, for example, include modals and modal alternatives (e.g., "ought to – it is essential that"), text connectives suitable for developing and structuring an argument (e.g., "first of all," "because of this," "this results in"), or language to contrast arguments and counterarguments (e.g., "on the other hand"; "While it is true that … it can also be argued that"). As pointed out by Spycher (2017), teachers often use model writing and show their students a mentor text (i.e., an example of good writing), but explicitly analyzing the language of that model text to facilitate classroom discussions is rarely adopted. Moving away from individual words and focusing on chunks, we propose collectively exploring a number of argumentative essays since this should help students see that language is not simply a set of rules but a range of choices that are made based on the content area and a writer's purpose. As described in EFL writing pedagogy (Hyland, 1990), this could help them understand the difference between "advertising" and "advertisement" or avoid inventing non-existent linking words such as "firstable." Building on our results and Wu et al.'s experimental study exploring the potential of model texts with Chinese EFL learners (2023), we advocate for implementing similar interventions in German and Swiss classrooms. Finally, we argue that this approach could also be a valuable asset in writing classes of other widespread L2s of the two educational systems under investigation (e.g., French, Spanish or Italian). Not only do the curricula of these subjects underline the importance of argumentative writing (Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein, 2014), but they also include word chunks such as *tout d'abord* (French for "first of all"), *a pesar de que* (Spanish for "in spite of"), or *a patto che* (Italian for "on condition that"). Given that many students in Swiss vocational schools struggle with writing (Konstantinidou et al., 2023) and because minority language children score significantly lower in German writing exams (Chudaske, 2012), scrutinizing model texts could become an essential strategy in L1 lessons in both countries, as this has been proven effective in Graham et al.'s meta-study on successful L1-writing interventions in other linguistic contexts (Graham et al., 2023).

Large Language Models (LLMs) such as ChatGPT also offer new possibilities of addressing grammatical and lexical issues in a differentiated and individualized way in EFL writing instruction. These tools can assist students in identifying mistakes in their texts and thereby contribute to notable gains in their performances (Escalante et al., 2023). With respect to the two features considered in our analysis, an LLM might highlight grammatical and lexical errors, enabling learners to correct their own work without extensive teacher feedback. Alternatively, AI can produce model texts or improved versions of an essay, which students can compare with their drafts to deepen their knowledge of argumentative collocations or grammatical tenses. However, while integrating LLMs in the classroom aligns with modern technological advancements, it is essential to balance this trend with curricular requirements. As stipulated in both countries, students must develop the ability to write proficiently without relying on AI. Nevertheless, it is undeniable that LLMs offer substantial support in individual practice and iterative writing processes, making them a very powerful tool in modern writing education that remains to be explored and investigated more thoroughly.

5.5 Conclusion, limitations and directions for future research

This study examined upper secondary school students' grammatical and lexical skills when writing argumentative essays over the course of 1 year. It built on prior research by using a more detailed analytic rubric, increasing the number of study participants and extending the time period to one entire school year. There are, however, three limitations that need to be considered. One of the biggest drawbacks was the selective operationalization of grammatical and lexical skills and the fact that we solely looked at *accuracy,* which differs from approaches taken by Read (2000). Secondly, stopping the counting process after seven errors resulted in a ceiling effect and more finely tuned scales at the top edge might have led to normal distributions in our data sets. Thirdly, while all texts in our study were written online to ensure consistency, we did not consider how the digital modality itself might have influenced students' writing performance. Previous research (cf. Pikhart et al., 2023) has highlighted cognitive differences between reading in print and on screen, which could similarly affect writing processes and the nature of learners' difficulties.

Future research could take a closer look at the different types of grammatical and lexical errors. While we did not distinguish between wrong nouns, wrong articles or wrong prepositions for example, sub-classifying the categories further would be insightful for the students and the teachers involved. In addition, a similar analysis could be undertaken when paying special attention to the structure and the content of the essays. In other words, it would be worth the effort to examine whether the number of *hooks* in introductions or the number of *concluding statements* in conclusions increases over time in both countries. Finally, targeted intervention studies could monitor whether carefully selected exercises or Large Language Models can help the students to improve their grammatical and lexical competencies more rapidly.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving human samples in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

FL: Formal analysis, Data curation, Writing – review & editing, Software, Methodology, Conceptualization, Investigation, Writing – original draft. RT: Software, Writing – review & editing, Methodology, Data curation, Conceptualization. JL: Data curation, Writing – review & editing, Methodology, Software. JM: Supervision, Writing – review & editing, Conceptualization, Funding acquisition, Project administration. TJ: Methodology, Writing – review & editing. SK:

Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the Swiss National Science Foundation (Grant No. 100019L162675) and the German Research Foundation (Grant No. KO1513/12-1). The funders had no role in the study design, data collection, analysis, interpretation, or manuscript writing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1605658/full#supplementary-material

References

Barkaoui, K. (2010). Explaining ESL essay holistic scores: a multilevel modeling approach. *Lang. Test.* 27, 515–535. doi: 10.1177/0265532210368717

Barkaoui, K., and Hadidi, A. (2020). *Assessing change in English second language writing performance*. New York: Routledge.

Biesenbach-Lucas, S. (2007). Students writing emails to faculty: an examination of E-politeness among native and non-native speakers of English. *Lang. Learn. Technol.* 11, 59–81. doi: 10125/44104

Bildungsdirektion des Kantons Zürich. (2017). *Lehrplan für die Volksschule des Kantons Zürich* [Curriculum]. Available online at: https://zh.lehrplan.ch/ (Accessed March 3, 2025).

Božić Lenard, D., Ferčec, I., and Liermann-Zeljak, Y. (2018). "Grammar or vocabulary - students' friends or foes?" in *Establishing predominance of English for specific purposes within adult English language teaching*. eds. N. Stojković and N. Burkšaitienė (Newcastle upon Tyne: Cambridge Scholars Publishing), 1–26.

Breland, H., and Jones, R. J. (1984). Perceptions of writing skills. *Written Commun.* 1, 101–109. doi: 10.1177/0741088384001001005

Brown, D. (1994). *Principles of language learning and teaching*. Englewood Cliffs, N.J: Prentice Hall Inc.

Chudaske, J. (2012). *Sprache, Migration und schulfachliche Leistung: Einfluss sprachlicher Kompetenz auf Lese-, Rechtschreib- und Mathematikleistungen*. Wiesbaden: VS Verlag.

Council of Europe (2001). *Common European framework of reference for languages*. Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Q.* 34, 213–238. doi: 10.2307/3587951

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Mod. Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

Cushing, S. T. (2019). "Assessment of writing" in *The encyclopedia of applied linguistics*. ed. C. A. Chapelle (Hoboken, NJ: Wiley), 1–7.

Darus, S., and Subramaniam, K. (2009). Error analysis of the written English essays of secondary school students in Malaysia: a case study. *Eur. J. Soc. Sci.* 8, 483–494.

De Smedt, F., Landrieu, Y., De Wever, B., and van Keer, H. (2022). Do cognitive processes and motives for argumentative writing converge in writer profiles? *J. Educ. Res.* 115, 258–270. doi: 10.1080/00220671.2022.2122020

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Educational Testing Service (2009). *The official guide to the TOEFL test*. New York: McGraw-Hill.

Erziehungsdepartement Basel-Stadt. (2013). *Lehrplan gymnasium* [curriculum]. Available online at: https://www.edubs.ch/unterricht/lehrplan/mittelschulen (Accessed March 3, 2025).

Escalante, J., Pack, A., and Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *Int. J. Educ. Technol. High. Educ.* 20. doi: 10.1186/s41239-023-00425-2

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Q.* 28, 414–420. doi: 10.2307/3587446

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., and Köller, O. (2019). Linking TOEFL iBT® writing rubrics to CEFR levels: cut scores and validity evidence from a standard setting study. *Assess. Writing* 43, 33–47. doi: 10.1016/j.asw.2019.100420

Fritz, E., and Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assess. Writing* 18, 173–181. doi: 10.1016/j.asw.2013.02.001

Graham, S., Kim, Y.-S., Cao, Y., Lee, J. W., Tate, T., Collins, P., et al. (2023). A meta-analysis of writing treatments for students in grades 6–12. *J. Educ. Psychol.* 115, 1004–1027. doi: 10.1037/edu0000819

Harmon, J. M., Hedrick, W. B., and Wood, K. D. (2005). Research on vocabulary instruction in the content areas: implications for struggling readers. *Read. Writ. Q.* 21, 261–280. doi: 10.1080/10573560590949377

Hascher, T. (2005). Emotionen im Schulalltag: Wirkungen und Regulationsformen. *Z. Pädag.* 51, 610–625. doi: 10.25656/01:4771

Henry, J. (2000). *Writing workplace cultures: An archaeology of professional writing*. Carbondale, IL: Southern Illinois University Press.

Higgs, T. V., and Clifford, R. (1982). "The push toward communication" in *Curriculum, competence, and the foreign language teacher*. ed. C. James (Skokie, IL: National Textbook Company), 55–78.

Hyland, F. (2003). Focusing on form: student engagement with teacher feedback. *System* 31, 217–230. doi: 10.1016/S0346-251X(03)00021-6

Hyland, K. (1990). A genre description of the argumentative essay. *RELC J.* 21, 66–78. doi: 10.1177/003368829002100105

Institute for Educational Quality Improvement [IQB]. (2021). *Sekundarstufe II – Englisch [educational resource]*. Available online at: https://www.https://www.iqb.hu-berlin.de/bista/UnterrichtSekII/englisch/ (Accessed March 3, 2025).

Jacobs, H. L. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

James, C. (1998). *Errors in language learning and use*. London: Routledge.

Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe: Theorie, Prozessgestaltung*, Empirie. Giessener Beiträge zur Fremdsprachendidaktik. Narr

Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., and Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing*, 48, 1–13. doi: 10.1016/j.jslw.2019.100700

Keller, S. D., Lohmann, J., Trüb, R., Fleckenstein, J., Meyer, J., Jansen, T., et al (2024). Language quality, content, structure: What analytic ratings tell us about EFL writing skills at upper secondary school level in Germany and Switzerland. *Journal of Second Language Writing*, 65, 101–129. doi: 10.1016/j.jslw.2024.101129

Kim, M. (2021). Considerations and challenges in longitudinal studies of lexical features in L2 writing. *Vocab. Learn. Instruct.* 10, 82–90. doi: 10.7820/vli.v10.2.kim

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Lang. Test.* 28, 509–541. doi: 10.1177/0265532211400860

Knoch, U., Rouhshad, A., and Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assess. Writing* 21, 1–17. doi: 10.1016/j.asw.2014.01.001

Köller, O., Fleckenstein, J., Meyer, J., Paeske, A. L., Krüger, M., Rupp, A. A., et al. (2019). Schreibkompetenzen im Fach Englisch in der gymnasialen Oberstufe. *Z. Erziehungswiss.* 22, 1281–1312. doi: 10.1007/s11618-019-00910-3

Konstantinidou, L., Madlener-Charpentier, K., Opacic, A., Gautschi, C., and Hoefele, J. (2023). Literacy in vocational education and training: scenario-based reading and writing education. *Read. Writ.* 36, 1025–1052. doi: 10.1007/s11145-022-10373-4

Lahuerta, A. C. (2018). Study of accuracy and grammatical complexity in EFL writing. *Int. J. English Stud.* 18, 71–89. doi: 10.6018/ijes/2018/1/258971

Landrieu, Y., De Smedt, F., Van Keer, H., and De Wever, B. (2022). Assessing the quality of argumentative texts: examining the general agreement between different rating procedures and exploring inferences of (dis)agreement cases. *Front. Educ.* 7:784261. doi: 10.3389/feduc.2022.784261

Lee, K. R. (2016). Diversity among NEST raters: how do new and experienced NESTs evaluate Korean English learners' essays? *Asia-Pac. Educ. Res.* 25, 549–558. doi: 10.1007/s40299-016-0281-6

Leki, I., and Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Q.* 28, 81–101. doi: 10.2307/3587199

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward. Language teaching publications series*. Hove, UK: Language Training Publications.

Linacre, J. M. (2018) *A user's guide to FACETS: Rasch-model computer programs* [software manual]. Available online at: https://www.winsteps.com/manuals.htm (accessed November 18, 2024).

Llach, M. P. A. (2011). *Lexical errors and accuracy in foreign language writing*. Bristol: Multilingual Matters.

Llach, M. P. A. (2017). Vocabulary teaching: insights from lexical errors. *TESOL Int. J.* 12, 63–74.

MacKenzie, I. (2013). *English, as a lingua franca*. London: Routledge.

Macqueen, S. (2012). *The emergence of patterns in second language writing: A Sociocognitive exploration of lexical trails (linguistic insights)*. Bern: Peter Lang Publishing Group.

McNamara, T. F. (Ed.) (1996). *Measuring second language performance. Applied linguistics and language study*. London: Longman.

Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein. (2014). *Fachanforderungen Englisch* [Curriculum]. Available online at: https://fachportal.lernnetz.de/sh/fachanforderungen/englisch.html (Accessed March 3, 2025).

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (2021) *Korrektur von schriftlichen Arbeiten im Fach Englisch (Sek. I)*. Available online at: https://www.standardsicherung.schulministerium.nrw.de/cms/zentrale-pruefungen-10/faecher/getfile.php?file=2660 (Accessed March 3, 2025).

Nuruzzaman, M., Shafiqul Islam, A. B. M., and Jahan Shuchi, I. (2018). An analysis of errors committed by Saudi non-English major students in the English paragraph writing: a study of comparisons. *Advanc. Lang. Literary Stud.* 9, 31–39. doi: 10.7575/aiac.alls.v.9n.1p.31

Pikhart, M., Klimova, B., Meunier, F., Ibarra, I., Suñer Muñoz, F., Zamborova, K., et al. (2023). A systematic review of the cognitive impact of digital media modalities on reading comprehension in L2. *Investig. Sobre Lectura* 18, 56–87. doi: 10.24310/isl.2.18.2023.16655

Polio, C., and Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *J. Second. Lang. Writ.* 26, 10–27. doi: 10.1016/j.jslw.2014.09.003

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rezaei, A. R., and Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assess. Writing* 15, 18–39. doi: 10.1016/j.asw.2010.01.003

Richards, J. C., and Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S. D., and Köller, O. Automated essay scoring at scale: a case study in Switzerland and Germany *ETS Res. Rep. Ser.* (2019) 2019 1–23 doi: 10.1002/ets2.12249 1

Schoonen, R., Snellings, P., Stevenson, M., and van Gelderen, V. (2009). "Towards a blueprint of the foreign language writer: the linguistic and cognitive demands of foreign language writing" in *Writing in foreign language: Contexts learning, teaching, and research*. ed. R. M. Manchón (Bristol: Multilingual Matters), 77–101.

Sermsook, K., Liamnimitr, J., and Pochakorn, R. (2017). An analysis of errors in written English sentences: a case study of Thai EFL students. *Engl. Lang. Teach.* 10, 101–110. doi: 10.5539/elt.v10n3p101

Song, B., and Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *J. Second. Lang. Writ.* 5, 163–182. doi: 10.1016/S1060-3743(96)90023-5

Spycher, P. (2017). *Scaffolding writing through the "teaching and learning cycle"*. San Francisco, CA: WestEd.

Storch, N., and Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *J. Engl. Acad. Purp.* 8, 207–223. doi: 10.1016/j.jeap.2009.03.001

Swan, M. (1988). *Practical English usage*. Oxford: Oxford University Press.

Thornbury, S. (1999). *How to teach grammar*. Harlow: Longman.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., and Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assess. Writing* 39, 50–63. doi: 10.1016/j.asw.2018.12.003

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C., Boldt, H., and Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: a pilot study. *TESOL Q.* 37, 345–354. doi: 10.2307/3588510

Wilcox, R. R. (2021). *Introduction to robust estimation and hypothesis testing*. Cambridge, MA: Academic Press.

Wolf, M. K., Oh, S., Wang, Y., and Tsutagawa, F. S. (2018). Young adolescent EFL students' writing skill development: insights from assessment data. *Lang. Assess. Q.* 15, 311–329. doi: 10.1080/15434303.2018.1531868

Wu, Z., Qie, J., and Wang, X. (2023). Using model texts as a type of feedback in EFL writing. *Front. Psychol.* 14, 1–12. doi: 10.3389/fpsyg.2023.1156553

Yoon, H. J. (2018). The development of ESL writing quality and lexical proficiency: suggestions for assessing writing achievement. *Lang. Assess. Q.* 15, 387–405. doi: 10.1080/15434303.2018.1536756

Zhan, H. (2015). Frequent errors in Chinese EFL learners' topic-based writings. *Engl. Lang. Teach.* 8, 72–81. doi: 10.5539/elt.v8n5p72