Assessing Writing 68 (2026) 101036

Contents lists available at ScienceDirect

Assessing Writing

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

Pursuing fair writing assessment: Halo effects in primary school foreign language writing in grade six

Ruth Trüb [a,*], Julian Lohmann [b], Jens Möller [c], Stefan D. Keller [d]

[a] University of Applied Sciences and Arts Northwestern Switzerland, School of Education, Institute for Primary Education, Bahnhofstrasse 6, 5210 Windisch, Switzerland
[b] Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany
[c] Kiel University, Institute for Psychology of Learning and Instruction, Olshausenstraße 75, 24118 Kiel, Germany
[d] Zurich University of Teacher Education, Lagerstrasse 2, 8090 Zürich, Switzerland

ARTICLE INFO

Keywords:
Writing assessment
Halo effects
Text characteristics
English as a foreign language (EFL)
Primary school

ABSTRACT

Assessing the writing competence of pupils learning English as a foreign language (EFL) at primary school is associated with specific challenges because of learners' limited language resources. This study investigates the extent to which characteristics of their texts trigger so-called halo effects. Halo effects are an assessment bias where the quality of one feature unintentionally influences the evaluation of other aspects. The study examines halo effects across nine aspects of text quality (communicative effect, level of detail, coherence, cohesion, complexity of syntax and grammar, correctness of syntax and grammar, vocabulary, orthography and punctuation), based on a random sample of narrative texts from a sixth-grade corpus. 200 pre-service teachers assessed four randomly assigned texts. Halo effects were calculated by comparison to expert ratings using multi-level regression analyses. Results show that orthography and vocabulary were the two main triggers of halo effects. Punctuation also triggered some halo effects, but to a smaller extent. The assessment of communicative effect, complexity and correctness of syntax and grammar was not determined by the corresponding text quality but dominated by other criteria. Results highlight the importance of being aware of halo effects when assessing young EFL learners' texts and emphasise the need for suitable training measures.

1. Introduction

Assessing children's communicative writing competence in a foreign language is associated with specific challenges due to pupils' limited language resources (Konrad et al., 2018). Spelling, for example, can have an impact on how well the reader understands the content, or missing punctuation can influence their perception of how well a text is structured (Trüb, 2022). This study is part of a larger research project that tries to shed light on these difficulties from an empirical perspective and explores training measures to reduce them.

While the acquisition of oral competences is often given priority in early years of foreign language instruction (Tono & Díez-Bedmar, 2014), curricula usually require learners to develop basic writing competences as well (e. g. Deutschschweizer

* Correspondence to: FHNW School of Education, Institute for Primary Education, Bahnhofstrasse 6, 5210 Windisch, Switzerland.
E-mail addresses: ruth.trueb@fhnw.ch (R. Trüb), lohmann@leibniz-ipn.de (J. Lohmann), jmoeller@ipl.uni-kiel.de (J. Möller), stefandaniel.keller@phzh.ch (S.D. Keller).

https://doi.org/10.1016/j.asw.2026.101036
Received 23 September 2025; Received in revised form 23 January 2026; Accepted 9 March 2026
Available online 23 March 2026

Erziehungsdirektoren-Konferenz, 2016). Typically, pupils learn how to write short texts from different genres such as text messages, letters, e-mails, factual texts, stories, recipes, weather forecasts or advertisements (Szabo, 2016a, 2016b). Initially, these texts are short and simple and gradually increase in length and complexity over the years (Hallet, 2016). Teachers use learners' texts to estimate their writing competence, give feedback, help them set new learning goals or to plan further lessons and writing tasks (Grotjahn & Kleppin, 2017; Hasselgreen & Caudwell, 2016).

Teachers and language testing professionals, however, face considerable challenges when they assess beginner-level learners' texts (Konrad et al., 2018). Studies show that most primary school EFL learners are capable of writing texts from various genres, but that the language they use is far from perfect (e. g. Trüb, 2022; Hasselgreen et al., 2012). When challenged by limited vocabulary resources, learners often use words from another language, substitute alternatives for unknown words ("hocky is my love sport") or simply omit certain words (Trüb, 2022, pp. 153). Sentences are sometimes incomplete and can show a considerable influence of the school language (Trüb, 2022). The learners' use of grammatical structures is not yet stable (the same structure is sometimes used correctly and sometimes incorrectly), and even strong learners still make basic orthographic and syntactical errors (Trüb, 2022). What at first glance might look like a lack of competence is a natural progression in language development as it can be observed in both first (L1) and second or foreign (L2) language acquisition (Péry-Woodley, 1991; Trüb, 2022). Such linguistic limitations can lead to a perception of text quality that is strongly shaped by salient features that attract the reader's attention (Trüb, 2022; Konrad et al., 2018). We know from research that certain text features influence how other text characteristics are assessed (Jansen et al., 2021; Vögelin et al., 2019). This type of bias is called halo effect. It describes a reader's tendency to mistakenly see unrelated or weakly related features as connected (Jacobs & Kozlowski, 1985). Studies examining whether and to what extent specific text characteristics trigger halo effects in primary school EFL writing assessment, however, are missing, and only few studies exist for older and more advanced EFL learners (e. g. Vögelin et al., 2019). Most existing studies investigated such halo effects with a small number of texts in which one aspect of text quality was modified, while all other text qualities remained unchanged. Although this research design has the advantage of isolating the effect of one specific text characteristic, it also lacks a certain generalisability regarding its representativeness for the target group of learners. Moreover, most studies focused on the influence of one specific text feature, and it is therefore unclear how halo effects present themselves when various text characteristics are considered jointly. Our study addressed these gaps by analysing halo effects in a primary school EFL context (grade six) and comprehensively across nine different aspects of text quality. Furthermore, we used a representative sample of authentic, unmodified texts from a young learner corpus to ensure a higher degree of generalisability. 200 pre-service primary school teachers each assessed four randomly assigned texts. Using multi-level regression analysis, their assessments were compared to validated benchmark ratings of intensively trained expert raters with extensive experience in primary school teaching and assessment.

The aim of the study was to identify text characteristics that trigger halo effects in primary school EFL writing assessment so that concrete measures of support can be taken. Knowledge about potential triggers of halo effects can help pre- and in-service teachers in their pursuit of fair writing assessment and support teacher educators in designing relevant and customised language assessment courses. Furthermore, a better understanding of halo effects in writing assessment can highlight important issues of judgment accuracy in writing research, provide guidance for large scale rater training and encourage further research in this area.

This article first presents the theoretical background of the study and the research questions, followed by a description of the research methods and results. Finally, it discusses the findings of the study in the light of previous research and draws conclusions for teacher education, teaching practice, theory and research.

2. Theoretical background

2.1. Challenges in assessing young EFL learners' writing competence

Research shows that language teachers express a great need for professional development in language assessment, and that the topic is often insufficiently covered in teacher training (Hasselgreen et al., 2004; Patekar, 2021; Vogt & Tsagari, 2014). A cross-national study among foreign language teachers in Europe (mainly from primary and secondary education) revealed that 73% of the teachers wished to receive additional training in assessing the productive language competences of speaking and writing (Vogt & Tsagari, 2014). From a survey among language testing professionals, we know that reliably differentiating between levels presents a particular difficulty when assessing beginner-level foreign language learners (Konrad et al., 2018). The most frequently mentioned challenge in responses to an open-ended question about assessing low-level learners was their limited language proficiency (Konrad et al., 2018, p. 26). Participants mentioned that the shortness of texts, limited orthographic correctness, punctuation errors, first language influence and limited vocabulary knowledge made the assessment difficult, and that coherence/cohesion was especially difficult to assess (Konrad et al., 2018). Some participants saw a danger in focusing too strongly on errors during the assessment or in expecting more accuracy than is warranted (Konrad et al., 2018). Some also expressed the difficulty of assessing all aspects fairly without being influenced by weaknesses in the texts (Konrad et al., 2018). Overall, 59% of the participants in the study experienced difficulties when assessing the writing performance of beginner-level foreign language learners (Konrad et al., 2018, p. 27).

Besides these qualitative studies, there are also two quantitative studies that examined the assessment of young EFL learners' writing skills. Zhu and Urhahne (2015) and Trüb et al. (2025) examined teachers' judgment accuracy (TJA) when assessing young EFL learners' texts. Zhu and Urhahne (2015) found a relative judgment accuracy of.77 between experts and teachers when they assessed texts of 5th-grade EFL learners on grammatical and orthographic correctness. Trüb et al. (2025) analysed TJA among pre-service teachers who assessed 6th-grade EFL learners' texts and found a relative TJA of $r = .34–.60$ for different assessment criteria. They found significant differences in how accurately the criteria were assessed. Vocabulary was assessed with the highest accuracy ($r = .60$),

followed by orthography, complexity and correctness of syntax and grammar, and punctuation ($r = .56–.49$) (Trüb et al., 2025). The criteria cohesion, level of detail, communicative effect and coherence (mainly pragmatic text features expressing the function, content and structure of a text) were assessed significantly less accurately than the other criteria (Trüb et al., 2025). These findings show that different text characteristics are assessed with different degrees of accuracy and that linguistic criteria such as vocabulary and orthography might be easier to assess accurately than pragmatic text qualities. Trüb et al. (2025) additionally found that pre-service teachers were stricter than experts on most criteria and that they assessed the texts with significantly more variability. This is highly relevant since primary school EFL learners are more anxious about learning English and have a lower perception of their competences if their performance is underestimated (Zhu & Urhahne, 2015). Moreover, it is well documented that self-efficacy is an important predictor of student performance (Pajares, 2003; Sun et al., 2021), also in EFL writing at primary school (Trüb, 2022), and therefore such findings are of paramount importance for teacher education and research. Not only should pre-service teachers receive guidance in assessing pragmatic text qualities, but they should also be given opportunities to calibrate their assessment with that of experts to avoid overly strict judgments that might negatively influence student learning. It also shows that research should not only investigate student assessment at a holistic level but also consider differences between assessment criteria.

Overall, we can see that the challenges in assessing young EFL learners' writing competence are manifold. The type of bias addressed in this article, halo effects, has so far received little attention in primary school EFL writing research. The following section will thus report findings from related contexts such as L1 learning and EFL learning at higher levels.

2.2. Halo effects in writing assessment

Research shows that halo effects are a common phenomenon in writing assessment. Much research on halo effects in writing assessment has been conducted in the context of rater-training and rater-mediated assessment (e. g. Ayoobiyan & Ahmadi, 2023; Engelhard, 1994), where the main purpose of the analyses is usually to adjust test scores for different types of bias or to identify raters who might need further training. Another branch of studies has investigated halo effects in teaching contexts with the aim of promoting teachers' assessment competence and supporting fair writing assessment in the classroom. Most of these studies focus on the specific influence of text characteristics on assessment quality and were conducted in L1 contexts or in EFL learning at upper secondary school.

In L1 learning, Lai et al. (2012, 2015) found evidence of halo effects when raters assessed multiple features of expository science texts written by middle school students. Their analyses suggested that not all text features are equally likely to trigger halo effects when one rater assesses all features (Lai et al., 2015). They found that text organisation was most affected by halo and the criterion of writing conventions (spelling, capitalisation, punctuation and grammar) was least affected (Lai et al., 2015). Rezaei and Lovorn (2010) used two manipulated texts at graduate level to examine the effect of rubrics in writing assessment. They found that the assessment of content was strongly influenced by the mechanical text quality, even when the raters used a rubric for the assessment. When the quality of grammar and spelling was high, the content was also rated positively, even though the text did not address the required content (Rezaei & Lovorn, 2010). Similarly, when the quality of grammar and spelling was poor, the content of the text was also rated low, even though the text adequately addressed the required content (Rezaei & Lovorn, 2010).

In EFL contexts, most research from this branch of studies was conducted at upper secondary-school level. Vögelin et al. (2018) examined the influence of the quality of lexis and orthography on teachers' feedback to learners. They found that the quality of lexis and orthography not only affected teachers' comments on lexical quality and orthography (correctly as intended) but also their comments on other criteria, thus indicating halo effects. The lexical quality had an influence on teachers' comments about grammar, and spelling on their comments about vocabulary, grammar and 'others' (such as rhetoric and style) (Vögelin et al., 2018). Vögelin et al. (2019) found that lexical quality also had an influence when teachers used rating scales to assess such texts. The results showed that texts with higher lexical quality received a more positive evaluation for overall text quality, frame of essay (introduction and conclusion) and grammar (Vögelin et al., 2019). In a third study, Vögelin et al. (2020) additionally found that the quality of text organisation influenced the assessment of all other criteria (vocabulary, task completion, support of arguments, grammar and spelling), thus indicating halo effects. These effects were stronger for texts with low overall text quality (Vögelin et al., 2020). Jansen et al. (2021) conducted a similar study and found that the quality of orthography led to halo effects, as it influenced the assessment of five other criteria (body of essay, support of arguments, grammar, vocabulary and task completion), but not the criterion frame of essay (introduction and conclusion). In all these EFL studies, the authors used an experimental design where they intentionally modified the text features under examination. They used two versions of the same texts, one with a high and the other with a low quality of the text feature. While this design has the advantage that differences in the assessment can be directly attributed to the text feature under examination, it also limits the generalisability of the results to the specific texts used in the study. If conclusions are to be drawn about a particular target group of learners as in our study, a larger and more representative sample of unmodified texts should be used for better generalisability (Möller et al., 2022). Lohmann, Lötscher et al. (2025) followed this approach and examined halo effects between language quality, structure and content in the assessment of argumentative essays written by EFL students in upper secondary school. They used randomly assigned texts from a large text corpus to obtain more robust and generalisable results. Their analysis showed that language quality triggered halo effects on structure and content, and content exerted halo effects on language quality and structure (Lohmann, Lötscher et al., 2025).

These studies show that text characteristics such as orthography, grammar, vocabulary, text organisation and content trigger halo effects in writing assessment. To our knowledge, however, no studies have been conducted to date to investigate halo effects in primary school EFL writing assessment. We therefore examined pre-service teachers' assessment of a randomised sample of narrative texts from a sixth-grade EFL text corpus. In contrast to most studies presented above, the analyses not only focused on halo effects triggered by

one specific assessment criterion, but aimed at a comprehensive analysis of nine different criteria, thus providing more detailed results about the influence of text characteristics on assessment quality than previously available.

2.3. Research questions and hypotheses

The following two research questions were addressed in the study:

RQ 1. : *Which aspects of text quality trigger halo effects in primary school EFL writing assessment?*

RQ 2. : *Which assessment criteria are affected by halo in primary school EFL writing assessment?*

Based on the studies summarised above, we expected that formal aspects such as vocabulary, orthography, grammar and punctuation were likely to trigger halo effects. In a primary school EFL context, problems at formal levels can be especially pronounced and might thus provide a strong lens through which the texts are viewed. By contrast, content- and structure-related criteria might be more prone to be affected by halo. However, since studies with more advanced EFL learners showed that formal aspects such as grammar, vocabulary, rhetorics/style or spelling can also be affected by halo (see 2.2), we expected that some language-related aspects could also be affected.

3. Research methods

3.1. Participants

Participants were 200 pre-service teachers studying Primary Education at a teacher college in [country]. They were, on average, 25.4 years old ($SD = 6.2$ years, ranging from 19 to 59 years), 70.5% were female and 29.5% male students. Most of them were in the second or third year of their studies, and they all had a minimum language proficiency of CEFR level B2 + (which is an official entry requirement for the second half of their English language teaching studies). According to an additional survey, 43% of the students had a C1 level (advanced), 9% a C2 level (proficient), and 2% had English as their first language. 42% of the students had not yet taught English at primary school, 42% a few lessons and 17% several lessons over a longer period. 77% had no prior experience in assessing English texts, 21% a little and 2% some more. They had all completed a first English language teaching (ELT) methodology course where they had been introduced to basic elements of foreign language classroom assessment such as different purposes and types of assessment, competency- and performance-based assessment and key quality criteria.

3.2. Assessment criteria and learner texts

For this study, we used assessment criteria from Trüb (2022) that had been specifically developed for assessing primary school EFL learners' texts in grade six. They covered the nine aspects communicative effect/creativity, level of detail, coherence, cohesion, complexity of syntax and grammar, correctness of syntax and grammar, vocabulary, orthography and punctuation (Trüb, 2022, pp. 102–103). Each aspect of text quality had been operationalised in the form of a 5-point assessment scale covering the CEFR beginner levels A1 and A2 (see Appendix A for the detailed descriptors of each scale). The scales had been developed based on genre analysis (Augst et al., 2007; Brinker et al., 2018; Hyland, 2004), CEFR descriptors relevant for the target age group (Szabo, 2016b) and already existing assessment grids for young learners (Brock, 2015; Hasselgreen et al., 2012). They had been thoroughly tested and revised with sample texts from two pilot studies following the *combined rater training and scale revision approach* by Harsch and Martin (2012), see Trüb (2022). The interrater reliability of the different scales ranged between .78 and .94 (ICC, two-way mixed, single measure, absolute agreement; see Trüb (2022), p. 121). They were thus considered suitable for this study.

The learner texts were randomly selected from a corpus of 322 narrative texts from the same study (Trüb, 2022). The texts had been written by 6th-grade primary school learners in their fourth year of learning English as a foreign language in [country]. Grade six in [country] is the end of level one of the International Standard Classification of Education (UNESCO Institute for Statistics, 2012), and learners are approximately 12–13 years old. The writing task was a picture story about a family meal to which the learners invented their own story ending (see Appendix B). 35 double-rated texts with a similar mean and criterion-specific distribution of scores to the entire corpus were randomly drawn from the corpus to ensure that the text sample adequately represented the entire corpus. The size of the text sample was determined to allow for an additional text-based analysis (not discussed in this article). After data collection, one text had to be excluded because it had an offensive content, which seems to have led to strong outliers in the participants' ratings. This led to a final sample of 34 texts used in the study. Appendix C provides a comparison of the text sample to the entire text corpus, including the mean ratings per criterion and the distribution of texts across language levels. Appendix D provides four text samples with the corresponding average expert ratings.

3.3. Expert ratings

For later data analysis, we used expert ratings as the "gold standard" against which pre-service teachers' ratings were compared. We relied on expert ratings that had already been used in an earlier study (Trüb, 2022) and evaluated them to ensure that they met the required quality standards. The two experts involved had both been primary school teachers for many years, had a high English language proficiency and extensive assessment experience. One of them was a specialist in English language teaching methodology and

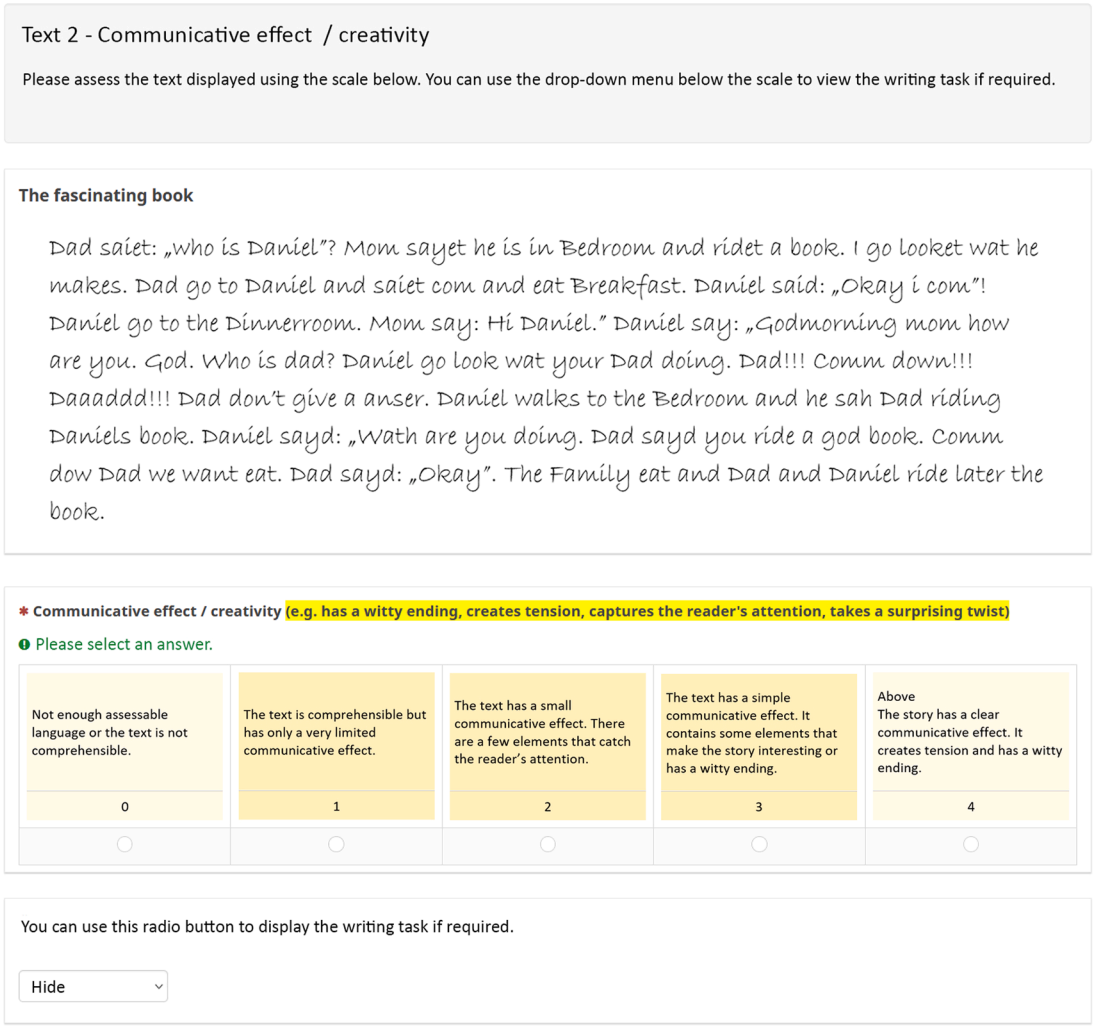


Fig. 1. Online assessment tool developed with the Open Source software LimeSurvey (LimeSurvey GmbH, 2021), *Note*. As can be seen in this sample text and those in Appendix D, formal problems at these levels can be especially pronounced and might thus trigger halo effects.

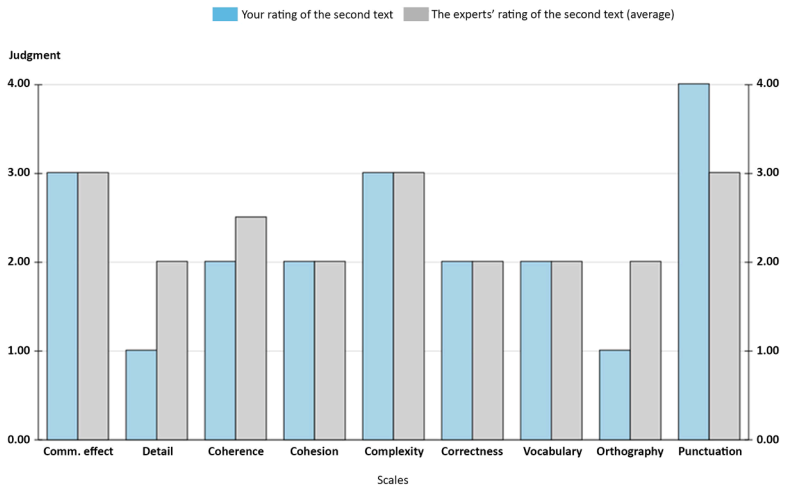


Fig. 2. Comparison with the experts' ratings (displayed after assessing all four texts).

worked as a lecturer at a teacher education college. They had undergone intensive rater training and reached an interrater reliability between .78 and .94 for the different assessment criteria (ICC, two-way mixed, single measure, absolute agreement; see Trüb, 2022, p. 121), indicating a high assessment reliability. Their ratings correlated by .70 with the English teachers' assessment of their learners' writing competence, indicating satisfactory criterion validity of the scores (Trüb, 2022, p. 126).

To evaluate the quality of the expert ratings and to explore whether they were affected by halo, we ran a many-facet Rasch analysis with the ratings of all texts from the corpus, using the software FACETS (Version 4.1.7; Linacre, 2025). This type of analysis is particularly suitable in scenarios where no benchmark ratings exist that can be used to assess rating quality. The Rasch model included the three facets examinee proficiency, raters and criteria. 5.17% absolute standardised residuals ≥ 2, and 0.83% ≥ 3 indicated a good model fit (Linacre, 2025). We analysed group-level statistics (criterion facet), individual-level statistics (rater facet) and conducted a bias interaction analysis between the criterion and rater facet to evaluate the presence of halo effects (Eckes, 2015; Myford & Wolfe, 2004).

Regarding the criterion facet, the analysis showed a criterion difficulty spread from −0.84 to 0.81 on the logit scale, with an average standard error of $SE = 0.04$. A homogeneity index of χ^2 (8) = 1488.1, $p < .001$ (indicating significant differences in difficulty between the criteria), a separation index of 18.63 (indicating more than 18 statistically distinct levels of criterion difficulty), and a high separation reliability index of .99 (indicating that raters could reliably distinguish between the criteria) gave no indication of group-level halo effects in the data set.

At the individual level (rater facet), the severity measures showed that the two experts were generally similarly strict, with a rater logit spread of only 0.12 compared to a 12.16-logit spread observed for examinee writing proficiency. The standard deviation of the rater facet ($SD = 0.09$) was substantially smaller than that of the examinee facet ($SD = 1.74$), suggesting that inter-rater variance accounted for only a small proportion of the variability in examinee performance. A rater separation reliability of 0.90 and a fixed-all-the-same chi-square test of χ^2 (1) = 20.4, $p < .001$ revealed a consistent and statistically significant difference in rater severity. However, given the very small magnitude of this difference relative to the variation in examinee proficiency, the difference is unlikely to be of practical relevance. The use of rater fit statistics to detect halo effects is complex since it depends on the specific measurement context (Eckes, 2015). Considering that criteria varied in difficulty (see criterion facet), pronounced differences in ratings are to be expected, meaning that halo effects are indicated by similar ratings across criteria and rater infit/outfit statistics significantly greater than one (Eckes, 2015). Individual rater infit/outfit statistics were close to one (ranging from 0.97 to 1.02), indicating an internally consistent use of the assessment scales and no indication of halo effects.

A bias/interaction analysis between raters and criteria showed that only 0.06% of the observed variance was explained by bias/interaction, while 56.3% of the variance was explained by the Rasch measures. 2 of 18 rater-criterion combinations showed a small but significant bias interaction. Rater 1 awarded higher scores than expected for communicative effect (0.18 logits, $SE = 0.07$, $t(485) = 2.60$, $p = .010$) and level of detail (0.14 logits, $SE = 0.07$, $t(485) = 1.99$, $p = .048$). As assessed by visual inspection of the corresponding bias diagram, none of the raters assigned highly similar ratings across the criteria, which would be an indication of halo effects. Overall, the analyses confirmed the high quality of the expert ratings, and we therefore concluded that they could be used as "gold standard" in this study.

Since each text in the text sample had been rated by two experts, a score had to be defined for those cases where they had not assigned the same score. Calculating the mean of the two ratings (e. g. $M = 3.5$) was discarded since the participants' ratings could only be integer values, and the corresponding difference would have automatically added unintended error variance to our calculations. We therefore used the score of the more balanced rater, as indicated by the raters' severity measures and infit/outfit statistics in the many-facet Rasch analysis.

3.4. Procedure

Data was collected in ELT seminars at a large teacher education college in [country]. In a 90-minute seminar session about EFL writing assessment at primary school, pre-service teachers were given a brief introduction about primary school EFL learners' writing competences and corresponding curricular requirements. They then reflected on possible assessment criteria and assessed four texts in an online assessment tool. At the end of the session, they discussed their insights and the relevance of this experience for classroom practice. To comply with research ethics, participants were informed about the aims of the study and their right to withdraw from participation at any time without any disadvantages and asked to give their informed consent.

The online assessment tool (see Fig. 1) randomly assigned four texts from the text sample to each participant. The number of texts per participant had been determined based on the time available in the seminar. The texts were presented in typewritten form (but not corrected for any errors) to ensure that the assessment was free from any unintended influence of the handwriting. First, the participants were introduced to the writing task (see Appendix B) and assessed the four texts holistically to familiarise themselves with the learners' overall writing competence. This was important because the great majority of participants had no prior experience in assessing English language learners' texts (see 3.1). They were then given an overview of the nine assessment criteria and assessed the same four texts with the scales described above (see Appendix A). In the end, they completed a questionnaire on demographic information, compared their own ratings with those of experts (see Fig. 2) and gave feedback on the online tool.

3.5. Data analysis

To measure halo effects, we followed the approach suggested by Henik and Tzelgov (1985) and West and Kenny (2011) and ran nine regression analyses with the experts' ratings of the nine assessment criteria as predictor variables and the participants' rating of

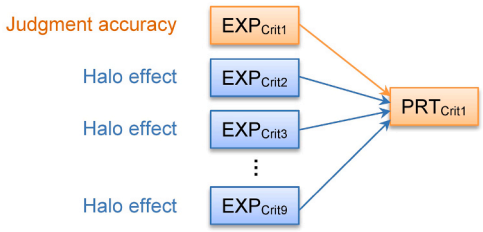


Fig. 3. Sample illustration of the regression model.

one criterion as dependent variable (see Fig. 3). This approach allows the identification of specific text quality features that trigger halo effects, as well as those that are affected by them.[1] All variables were z-standardised before they were included in the model. As described above, the expert ratings were considered as "gold standard" representing actual text quality. The expert rating of the criterion corresponding with the outcome variable (orange in the example) was expected to significantly predict the participants' rating. This prediction represents the pre-service teachers' judgment accuracy when controlled for the influence of all other criteria.

The expert ratings of the non-corresponding criteria (blue in the example) should ideally have no influence on the participants' ratings. For example, the quality of vocabulary should not influence the pre-service teachers' rating of the criterion punctuation. If there is a significant effect of a non-corresponding criterion on the pre-service teachers' rating, this would indicate a halo effect. By including the expert ratings of all nine criteria in the same model, the inter-correlations between them are accounted for (Henik & Tzelgov, 1985).

Since our data had a multi-level structure, we calculated multi-level regression analyses with $n = 772$ judgments on level 1 and $n = 200$ pre-service teachers on level 2. We used random intercepts models to allow the outcome variable to vary between participants and tested the data for linearity, normality of residuals, homoscedasticity and multicollinearity. The data met all requirements, as assessed by visual inspection of scatterplots, Q-Q-plots and VIF calculations. To account for multiple testing, we applied a Bonferroni-Holm correction (Holm, 1979).

To compare the corrected judgment accuracy in our models with the direct bivariate correlation between experts and pre-service teachers (also called rank component of teacher judgment accuracy, see Lohmann, Machts et al., 2025; Helmke & Schrader, 1987) we also calculated Pearson's correlation coefficients for each pair of corresponding ratings. The interested reader is referred to Trüb et al. (2025) for a detailed analysis of the different components of TJA for this study.

4. Results

Table 1 shows the results of the analyses. The grey cells in the table contain the regression coefficients of the corresponding criteria and, for reasons of comparison, the uncorrected bivariate correlations between the experts' and pre-service teachers' ratings (= rank component of TJA). The white table cells contain the regression coefficients of the non-corresponding criteria, with significant results indicating halo effects. Positive coefficients indicate a positive relationship, with higher (lower) expert ratings corresponding to higher (lower) ratings from the participants.

4.1. Aspects of text quality triggering halo effects (RQ 1)

As we can see in the white table cells, six out of nine criteria exert a significant influence on at least one other criterion. Vocabulary and orthography are the strongest triggers of halo effects: vocabulary affects the assessment of all other criteria except for communicative effect (β =.15–.32), and orthography affects the assessment of all other criteria except for coherence (β =.13–.28). Punctuation leads to bias in the assessment of the four criteria coherence, complexity of syntax and grammar, correctness of syntax and grammar and orthography (β =.10–.12). Coherence significantly influences the assessment of the communicative effect (β =.14), cohesion the correctness of syntax and grammar (β =.16), and complexity of syntax and grammar the range of vocabulary (β =.22).

4.2. Assessment criteria being affected by halo (RQ 2)

Overall, we can see that all assessment criteria are affected by halo. Most criteria are influenced by two other aspects of text quality, except for complexity and correctness of syntax and grammar, which are influenced by three and four other aspects respectively. The participants' ratings of six criteria are still significantly predicted by the corresponding text quality, even though there is some

[1] An anonymous reviewer noted that Structural Equation Modeling (SEM) could be used as an alternative modeling approach. We agree that SEM has considerable potential for research on teacher judgments (see Lohmann, Machts et al., 2025) and could also be applied to the dataset analyzed in the present study. However, our research questions focus on estimating the same set of predictor effects for each outcome, treating all variables as observed, rather than on modeling latent constructs, or comparing effects across equations. In addition, given the clustered data structure, a corresponding joint multilevel SEM would substantially increase model complexity and may yield less stable estimates with our sample size. We therefore rely on more parsimonious multilevel random-intercept regression models.

Table 1
Halo effects: multi-level regression coefficients β (SE_z) [a] and correlation coefficients (r, in the diagonal) [b].

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Pre-service teachers' ratings Expert ratings	Communica-tive effect	Level of detail	Coherence	Cohesion	Complexity syntax and grammar	Correctness syntax and grammar	Vocabulary	Orthography	Punctuation
Communicative effect	.01 (0.05) **.39** ***	.02 (0.05)	.05 (0.05)	-.13 (0.05)	-.03 (0.05)	-.08 (0.04)	-.00 (0.04)	-.10 (0.04)	-.07 (0.05)
Level of detail	.10 (0.05)	**.21 (0.05)** *** **.39** ***	.11 (0.04)	.01 (0.05)	-.04 (0.04)	.02 (0.04)	-.01 (0.04)	.02 (0.04)	-.00 (0.04)
Coherence	**.14 (0.04)** **	.11 (0.04)	**.13 (0.03)** ** **.34** ***	.09 (0.04)	-.01 (0.03)	-.04 (0.03)	-.01 (0.03)	-.07 (0.03)	.04 (0.03)
Cohesion	.04 (0.04)	.09 (0.04)	.02 (0.04)	**.28 (0.04)** *** **.40** ***	.11 (0.04)	**.16 (0.04)** **	.07 (0.04)	.06 (0.04)	-.02 (0.04)
Complexity syntax and grammar	.16 (0.06)	.04 (0.05)	.13 (0.05)	.07 (0.05)	.16 (0.05) **.50** ***	.16 (0.05)	**.22 (0.05)** ***	.07 (0.05)	.13 (0.05)
Correctness syntax and grammar	-.03 (0.05)	-.06 (0.05)	.02 (0.05)	.00 (0.05)	.02 (0.05)	.10 (0.05) **.50** ***	-.12 (0.04)	.13 (0.04)	.05 (0.05)
Vocabulary	.14 (0.05)	**.17 (0.05)** *	**.15 (0.05)** *	**.23 (0.05)** ***	**.32 (0.04)** ***	**.24 (0.04)** ***	**.40 (0.04)** *** **.60** ***	**.26 (0.04)** ***	**.20 (0.04)** ***
Orthography	**.14 (0.04)** *	**.14 (0.04)** *	.12 (0.04)	**.21 (0.04)** ***	**.22 (0.04)** ***	**.27 (0.03)** ***	**.28 (0.03)** ***	**.39 (0.03)** *** **.56** ***	**.13 (0.04)** *
Punctuation	.09 (0.03)	.09 (0.03)	**.10 (0.03)** *	-.00 (0.03)	**.12 (0.03)** **	**.11 (0.03)** *	.06 (0.03)	**.11 (0.03)** **	**.45 (0.03)** *** **.49** ***

Note. $n = 772$ judgments per assessment criterion; [a] β-coefficient with standardised standard error; * $p < .05$, ** $p < .01$, *** $p < .001$ with a Bonferroni-Holm correction for multiple testing calculated with the online calculator by Hemmerich (2016); [b] correlation between experts' and pre-service teachers' ratings for each criterion; * $p < .05$, ** $p < .01$, *** $p < .001$ (grey table cells). Significant results printed in **bold**. See Appendix E for the model fit metrics and the unique variance explained by the predictors.

influence from other criteria. There are, however, three criteria whose assessment is overruled by non-corresponding text quality, namely communicative effect, complexity of syntax and grammar and correctness of syntax and grammar (as indicated by non-significant beta coefficients in the grey table cells). Even though their uncorrected bivariate correlations indicate medium to high judgment accuracy ($r = .39$ and.50), this relationship is conditioned by non-corresponding aspects of text quality: communicative effect is determined by coherence (β =.14) and orthography (β =.14); complexity of syntax and grammar by vocabulary (β =.32), orthography (β =.22) and punctuation (β =.12); and correctness of syntax and grammar by orthography (β =.27), vocabulary (β =.24), cohesion (β =.16) and punctuation (β =.11). This indicates that participants do not assess the quality of communicative effect, complexity of syntax and grammar and correctness of syntax and grammar in isolation but are strongly influenced in their assessment by non-corresponding criteria.

5. Discussion

This study aimed at investigating halo effects in primary school EFL writing assessment in grade six. It extends existing research by examining halo effects comprehensively across nine different assessment criteria and by basing the analyses on a random sample of texts from an EFL learner corpus. Moreover, to our knowledge, it is the first study to investigate halo effects in the specific context of primary school EFL writing.

5.1. Discussion and implications of the findings

Our expectations were largely confirmed by the results. The analyses showed that vocabulary and orthography trigger the most and strongest halo effects among all nine examined criteria, with vocabulary affecting the assessment of all other criteria except for communicative effect, and orthography influencing the assessment of all other criteria except for coherence. Punctuation influenced four other criteria, but the coefficients were smaller than those of vocabulary and orthography, thus indicating weaker halo effects in comparison. The study therefore emphasises the dominant role of vocabulary and spelling in triggering halo effects in primary school EFL writing assessment. It extends existing research (e. g. Jansen et al., 2021; Vögelin et al., 2019) by highlighting the relative contribution of vocabulary and orthography to triggering halo effects in comparison with other text features, whose influence is smaller. The findings also align with observations made by language testing experts who reported that the quality of orthography, punctuation and vocabulary in low-level learners' texts made the assessment difficult (Konrad et al., 2018). These three aspects of text quality, thus, require special attention in primary school EFL writing assessment. It appears paramount that pre- and in-service teachers be made aware of the potential bias they can cause. Since learners do not develop all features of writing simultaneously and the quality of individual text features can vary greatly within the same text (Péry-Woodley, 1991; Trüb, 2022), it is essential to assess each aspect of text quality as independently as possible from other text features. Awareness raising, joint discussion of texts, comparison with expert ratings, analysing annotated text samples or other types of training could support teachers in their pursuit of fair and unbiased writing assessment. In teaching practice, teachers could ask individual learners to read their texts to them if comprehension is impaired or if orthography and punctuation are likely to lead to bias in the assessment. They could also teach them how to revise specific features of their texts (Griva & Chostelidou, 2013). In general, it appears central that teachers develop a natural habit of looking beyond orthographic errors, vocabulary range and punctuation when assessing primary school EFL learners' texts to ensure fair writing assessment.

In addition to these main results, we found that coherence affected the assessment of the communicative effect, cohesion the assessment of correctness of syntax and grammar, and complexity of syntax and grammar the assessment of vocabulary. These findings are slightly more difficult to interpret. There seem to be, however, some plausible explanations. If comprehension, for example, is impaired because there are gaps in the storyline, key information is missing or because it is unclear who is speaking (Trüb, 2022), this could also affect the perception of the communicative effect of a text. If a text is linguistically well linked and there are many connectors, it seems plausible that teachers might also tend to assign high ratings for correctness of syntax and grammar. Or, since English grammar is often lexically determined (Lewis, 2005), it may be difficult to assess the two aspects separately. It might be advisable, therefore, to help teachers recognise differences between these constructs and how they can be distinguished in the texts, for example by analysing sample texts and providing concrete examples.

The study also confirmed our expectation that content-, structure- and language-related criteria could be affected by halo. We found that all criteria were influenced by at least two other aspects of text quality. Contrary to our expectations, however, and despite medium and high bi-variate correlations, we found that the assessment of three criteria (communicative effect, complexity and correctness of syntax and grammar) was no longer determined by the corresponding text quality when the other eight criteria were included in the analysis. This indicates that these aspects were not assessed in isolation, but that their assessment was dominated by other aspects of text quality. It is possible that the three criteria may be difficult to identify in a text, that the assessment scales are difficult to apply for untrained examiners, or that the participants did not know what evidence to use to assess these criteria. Teacher training could address such difficulties by making teachers aware of how primary school EFL learners create a communicative effect or what they can expect from them in terms of grammar and syntax. Trüb (2022), for example, found that primary school EFL learners in grade six use various means such as tension and relief, unexpected elements, captivating direct speech, words of emphasis, distinctive use of punctuation, emotions, humour or interaction with the reader to make their texts interesting, and Lindgren and Stevenson (2013) and Griva et al. (2009) examined the various strategies 11- and 12-year-old learners use to overcome grammatical or syntactical obstacles caused by limited L2 resources (e. g. translation from L1, applying L1 rules to the foreign language, substitution or omission of elements). Teacher training and professional development courses could provide opportunities to analyse sample texts for such

elements so that teachers can use this knowledge for assessment and feedback. Moreover, they could jointly discuss and assess texts and compare their assessment with that of experts. It seems crucial that teachers should be well acquainted with the specificities of learners' writing competence at this level so that they can assess text quality fairly.

Besides contributing to teacher education and teaching practice, the study adds to our theoretical understanding of primary school EFL writing assessment. It provides a deeper understanding of halo effects across different criteria. To systematise factors influencing TJA, Südkamp et al. (2012) presented a model comprising the four categories teacher characteristics (e. g. teaching expertise), student characteristics (e. g. gender, ethnicity), judgment characteristics (e. g. holistic vs. analytic rating) and test characteristics (e. g. task difficulty). Our study shows that in the context of writing assessment such a model should additionally include text characteristics as a fifth category. Wolfe et al. (2016) refer to this category as *response content*, including, for example, a text's visual appearance, its textual quality or the quality of its content. In our case, it was the quality of different text features that influenced how accurately the texts were assessed.

Lastly, our study highlights how important it is that research on writing assessment should take the role of different aspects of text quality into account. It is often assumed that different text characteristics can be treated equally. We have shown, however, that different criteria are assessed with different degrees of accuracy (Trüb et al., 2025) and that they have an unequal potential for triggering halo effects.

5.2. Limitations

While this study provides important insights about halo effects in EFL writing assessment in grade six, there are also some limitations that need to be acknowledged. First, it is important to note that any assessment of written work is inherently subjective. While great efforts were made to ensure the reliability and validity of the experts' ratings, a certain subjectivity of even this gold standard may remain. Secondly, the study only relied on quantitative data and did not examine the assessment process. We are therefore unable to draw any conclusions about the reasons why certain text characteristics trigger halo effects. Thirdly, our study employed a relatively large set of assessment criteria, whereas in teaching practice usually a smaller number of assessment criteria is used. Based on our study design, we are unable to determine whether using fewer assessment criteria would prevent halo effects. Fourthly, we acknowledge that halo effects were calculated based on texts from only one text genre. Replication studies with texts from different genres could increase the generalisability of the results. Fifthly, the study relied solely on pre-service teachers, which limits the generalisability of the results, since more experienced teachers might exhibit different assessment behaviour.

5.3. Directions for future research

Future studies could contribute to our knowledge about halo effects and EFL writing assessment in several ways. First, it appears important to know what measures can be taken to reduce halo effects in primary school EFL writing assessment. Future studies could investigate whether and to what extent different training options or awareness-raising activities are able reduce halo effects in EFL writing assessment. This could support both teachers and language testing professionals in their pursuit of fair writing assessment. Secondly, we do not know why certain aspects of text quality trigger halo effects and why communicative effect, complexity and correctness of syntax and grammar are not determined by the corresponding text quality criteria. A close examination of assessment processes, for example by using eye-tracking technology, think-aloud protocols or stimulated recall interviews, might be able to shed light on these questions and contribute to a better understanding of how such assessment processes can be improved. Thirdly, it appears crucial to explore the role of artificial intelligence (AI) in primary school EFL writing assessment. With the widespread availability of AI, it is highly relevant that future studies examine the quality of AI-based writing assessments and investigate whether and to what extent they are affected by halo. Lastly, it appears central to acknowledge that education, including writing assessment, varies greatly in different educational and cultural environments. Replication studies could therefore provide further insights from different educational and cultural contexts.

5.4. Conclusion

This study investigated the role of halo effects in primary school EFL writing assessment in grade six. It extended existing research by using a randomised sample of texts from an EFL young learner corpus and considering nine different aspects of text quality jointly. It thus provided more detailed and generalisable results about halo effects in EFL writing assessment than were previously available. Furthermore, it extended previous research to the specific context of EFL writing at primary school.

The results showed that the most and strongest halo effects are triggered by orthography and vocabulary. Punctuation, complexity of syntax and grammar, cohesion and coherence also trigger halo effects, but to a considerably smaller extent. All nine assessment criteria were affected by halo, with three criteria even being dominated by other aspects of text quality.

These findings provide important information for classroom practice, teacher education, theory and research. First, they inform teachers and examiners about potential halo effects in primary school EFL writing assessment, thus putting them in a position to take corrective action. Moreover, the results support teacher educators when selecting relevant curricular content for EFL writing assessment and support them in helping pre-service and in-service teachers develop the necessary assessment competence to assess learners' writing performances reliably and fairly. The study by Trüb et al. (2025) already showed that content-related and structure-related aspects of text quality are assessed significantly less accurately than linguistic aspects, and that pre-service teachers assess primary school EFL learners' texts more strictly and with greater variability than experts. Our study adds to these findings by

highlighting the importance of being aware of halo effects. Assessment tools such as the one used for this study (see 3.4) could be developed further for the specific use in teacher education, so that pre-service and in-service teachers can practise assessing the writing competence of primary school EFL learners and be made aware of the specific challenges associated with it. This could support their assessment competence and contribute to high-quality assessment in everyday teaching practice. Moreover, if teachers feel confident about assessing primary school EFL learners' texts, they will use more communicative writing tasks from an early age, thus contributing to a wider use of communicative and competency-based language teaching. The study also contributes to theory building by providing a deeper understanding of halo effects across different criteria and of factors influencing teachers' accuracy of judgment. Finally, it contributes to research by emphasising the necessity of considering the role of different text characteristics when examining writing assessment.

CRediT authorship contribution statement

Ruth Trüb: conceptualisation, methodology, software, investigation, data curation, formal analysis, writing-original draft, writing-review&editing, project administration, funding acquisition. **Julian Lohmann:** conceptualisation, methodology, formal analysis, writing- review & editing. **Jens Möller:** conceptualisation, methodology, writing- review & editing, funding acquisition. **Stefan D. Keller:** conceptualisation, methodology, writing- review & editing, funding acquisition.

Funding

This work was supported by the Swiss National Science Foundation [197968], the German Research Foundation [315271436], the University of Applied Sciences and Arts Northwestern Switzerland and SwissUniversities.

Declarations of interest

The authors have nothing to declare.

Acknowledgements

We would like to thank the reviewers for their insightful comments and Marion Richner for proof-reading the article. S. D. G. – Soli Deo Gloria

Appendix A

Table A. 1
Analytic assessment criteria from Trüb (2022), pp. 102–103), based on Brock (2015), Hasselgreen et al. (2012) and Szabo (2016b)

	0 **Below**	**1** **Approx. level A1.1**	**2** **Approx. level A1.2**	**3** **Approx. level A2.1**	**4** **Above**
Communicative effect / creativity E. g. has a witty ending, creates tension, captures the reader's attention, takes a surprising twist	Not enough assessable language or the text is not comprehensible.	The text is comprehensible but has only a very limited communicative effect.	The text has a small communicative effect. There are a few elements that catch the reader's attention.	The text has a simple communicative effect. It contains some elements that make the story interesting or has a witty ending.	Above The story has a clear communicative effect. It creates tension and has a witty ending.
Level of detail E. g. detailed descriptions, reasons/ explanations, emotions	Not enough assessable language.	Describes the scene, people and actions without much detail.	Describes the scene, people and actions with a few details.	Describes the scene, people and actions with some details.	Above Describes the scene, people and actions in detail.
Coherence (logical organisation) Setting the scene, development/ complication, resolution, story ending	Text almost completely incoherent or not enough assessable language.	The writer does not fully succeed in developing a storyline. The text is incomplete and sometimes unclear.	It is evident that the writer is trying to tell a story. There is a very simple storyline with some gaps or incoherences.	The writer succeeds in telling a story with a simple and mostly coherent storyline. There are only few gaps or incoherences.	Above The storyline is coherent and contains all elements of a simple narrative structure.
Cohesion (linguistic connectivity) Repetition of key words, connectors (and, but,	No linking of words and phrases, only isolated chunks of language or not enough	Only few and very basic cohesive devices or reference words are used to link words and sentences (e. g. and).	A small number of very basic cohesive devices and reference words (e. g. and, this, it) are used to link words and	Some simple cohesive devices and reference words (e. g. and, but, because, then, or, he, she, they) are used to	Above A larger variety and amount of cohesive devices and reference words are used to link

(*continued on next page*)

Table A. 1 (*continued*)

	0 Below	1 Approx. level A1.1	2 Approx. level A1.2	3 Approx. level A2.1	4 Above
or, because, then,…), pronouns (he, she, they, it,…), demonstratives (this, that, there,…), comparatives (same, another, more…)	assessable language.	The text mainly consists of isolated phrases and sentences.	sentences. The text contains some isolated phrases and sentences.	link words and sentences. No or only few isolated phrases and sentences. Mostly linguistically well-linked text.	words, sentences and text passages. Linguistically well-linked text.
Complexity of syntax and grammar	Not enough assessable language.	Uses very simple and partly incomplete sentences and very simple grammatical structures, mainly in formulaic expressions.	Uses simple sentences (e. g. one-clause sentences) and simple grammatical structures (e. g. plural '-s'), often in formulaic expressions.	Uses a mixture of simple and more complex sentences and grammatical structures (e. g. simple subordinate clauses, 3rd person '-s', past or future forms, reporting clause for direct speech).	Above Uses more complex and varied sentences and grammatical structures.
Correctness of syntax and grammar	Not enough assessable language.	Only very limited control and many inconsistencies. Some reduction or omission of elements.	Only limited control and many inconsistencies.	Uses a few simple structures correctly but still systematically makes basic errors.	Above Uses some simple structures correctly but still makes basic errors.
Vocabulary range	Most of the text written in another language or not enough assessable language.	Basic chunks and limited vocabulary, some words and/or phrases in another language.	Basic vocabulary (e. g. simple home and family vocabulary, simple verbs, few adjectives), enough vocabulary to write the text mainly in English (no or only few words in another language).	Mainly simple but also some specific and/or varied vocabulary that allows a slight elaboration of the text.	Above Wider range of vocabulary that allows a clear elaboration of the text.
Orthography	Text hardly comprehensible or not enough assessable language.	Only very limited control and many inconsistencies. Comprehension sometimes impaired.	Only limited control and many inconsistencies, but the text is comprehensible.	Short and very common words are written with reasonable phonetic accuracy. Still makes basic errors.	Above Common words are written with reasonable phonetic accuracy. Still makes errors.
Punctuation	No or almost no use of punctuation.	Limited use of basic punctuation (e. g. full stops, question marks), sometimes incorrect or missing.	Mostly correct or correct use of basic punctuation (e. g. full stops, question marks), only occasionally missing OR Use of some further elements but basic punctuation sometimes incorrect or missing.	Correct use of basic punctuation. Use of some further elements, e. g. exclamation mark, colon or commas for a series of words or phrases, not necessarily correct. OR Use of some difficult elements but basic punctuation sometimes incorrect or missing.	Above Correct use of common punctuation. Use of some difficult elements, e. g. commas for dependent clauses or quotation marks, not necessarily correct.

Appendix B

Write a story

Look at the pictures. What happened at this family meal last week? What did the characters say and do? How did the story end?
Schau dir die Bilder an. Was geschah in dieser Familie letzte Woche beim Essen? Was sagten die verschiedenen Personen und was machten sie? Wie ging die Geschichte zu Ende?

On the basis of these pictures, write a coherent story with a witty ending. Include enough details so that the story becomes vivid and interesting for the reader.
Write the story in the past tense.
Schreibe auf der Grundlage dieser Bilder eine zusammenhängende Geschichte mit einem witzigen Ende. Schreibe so detailliert, dass die Geschichte für den Leser lebendig und interessant wird. Schreibe die Geschichte in der Vergangenheitsform.

The fascinating book

Das spannende Buch

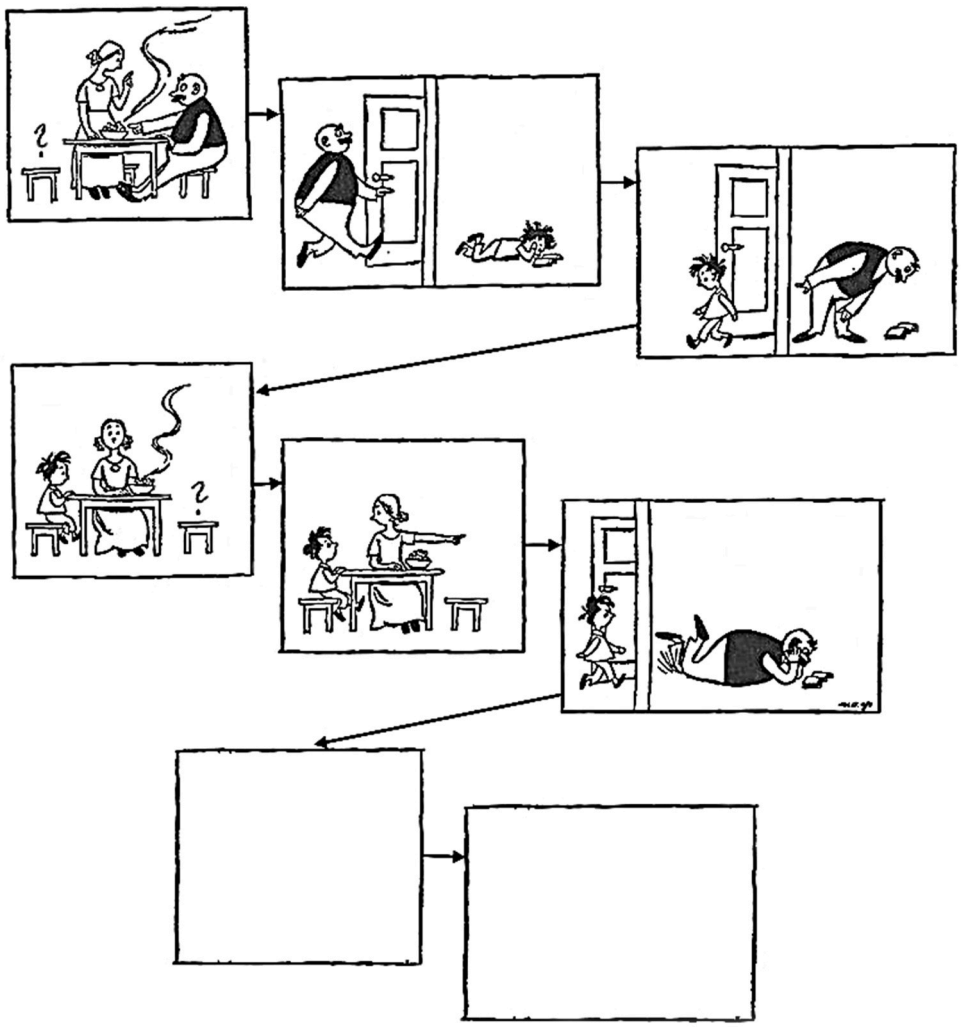
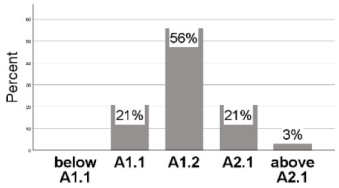
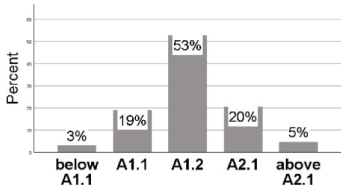


Fig. B. 1. Writing task "The fascinating book" from Trüb (2022, p. 110)

Appendix C

Table C. 1
Comparison of the selected text sample to the entire text corpus

	Text sample (n = 34) *M* (*SD*)	Corpus (n = 322) *M* (*SD*)
Mean rating[a]	**2.33 (0.59)**	**2.35 (0.70)**
Communicative effect / creativity	2.04 (0.64)	1.98 (0.82)
Level of detail	2.16 (0.87)	2.14 (1.00)
Coherence	2.21 (0.68)	2.21 (0.81)
Cohesion	2.56 (0.96)	2.73 (0.93)
Complexity syntax and grammar	2.53 (0.80)	2.60 (0.95)
Correctness syntax and grammar	2.18 (1.03)	2.32 (1.04)
Vocabulary	2.19 (0.83)	2.21 (0.88)
Orthography	2.56 (1.09)	2.43 (1.09)
Punctuation	2.57 (0.92)	2.59 (1.00)
Distribution across CEFR levels		

Note. [a] Unweighted average of the nine analytic criteria

Appendix D

Text samples from Trüb (2022, pp. 153–160)

1) The fascinating book (level A1.1)

   The fater and the mater food, dont the Kid ried a Book. The fater say: „come" The Kid came, the fater ried th Book, the mater say the Kid „going and bring de fater „Yes mom" say the Kid. The Kid coms not beak. „Mom comon its cool", the mater coing and then read the Book.

2) The fascinating book (level A1.2)

   „How is Jacob?" „He is in hes room, and read!", sayed mom to Vather: „Jacob, Jacob! Common, we are eating!" One moment it was still, then mom sayed vather: „Go and look!" Then the vather goes and seed Jacob is reading a book. „Jacob, comon we eating now!" Then Jacob gos eating and the vather is looked in the book. 10 min. laiters: „Jacob, look how is vather!", and then Jecon goed in hes room, and sead vather is reading the book. At the end of the story nofing will eating all will reading the book.

3) The fascinating book (level A2.1)

Last Week Family Müller wond eat lunch butt Max the son were not by the Table. Becors hes reading a fascinating book. The Dad say to the Mum: „Wer is Max?" the Mum meand: „Max is autside reading a book, go and get him!" The Dad of Max is going autside and say to him: „Max the Lunch is finish go to table!" Max is going inside and sit down but his dad have a look to the book. Now Max is sitting on the table but his Dad is not thear! His Mum told him: „Max have a look wer your Dad is!" Max is going autside and see his Dad reads in his book! He told him that the lunch is going cold but then he lied down too his Dad and read in the book with him!

4) The fascinating book (level A2.2)

It wose six o'clock. The Mother said: „Hans where is Lisa? Could you pleas get her here?" The Dad said okay. Hans wose walking threw the living room, he opend the door and said: „Lisa where you doing here?" Lisa said: „I wose reading this book!", Dad wosent very suprised Lisa loves to read. The fahter said with a deep voice: „enough reading its dinner time." Lisa said okay and she went to the dinig room, but dad wosent there. The Mother said to Lisa: „Lisa could you bring your dad here?" Lisa went there and saw what her Father wose doing there: „Dad! What are you doing here?" "Ahmm nothing!" „No dad i can see that you are reading my book about Barbies!" „But it is very intrestin!" „Aghh dad come we have to eat!" „Okay i am coming sweeti!" the two got to the eating table and Lisa saw that her dad wose eating and reading the book! She had to laugh very hard!

Table D. 1
Average expert ratings of text samples 1–4, from Trüb (2022, p. 292)

Criterion	Rating[a]			
	Text 1	Text 2	Text 3	Text 4
Story overall score[b]	**1.20**	**2.05**	**3.00**	**3.45**
Communicative effect / creativity	1.00	2.50	2.50	3.50
Level of detail	1.00	2.00	3.00	4.00
Coherence	1.00	1.50	3.50	3.00
Cohesion	1.00	2.50	3.50	2.50
Complexity of syntax and grammar	1.00	1.50	3.00	4.00
Correctness of syntax and grammar	1.00	1.50	3.00	4.00
Vocabulary	1.00	2.00	3.00	3.50
Orthography	1.00	2.00	2.50	2.50
Punctuation	3.00	3.00	3.00	4.00

Note. [a] Mean rating of the two raters. [b] Overall scores calculated by the means of the subcategories *task completion* (communicative effect/creativity, level of detail), *text structure and cohesion* (coherence, cohesion), *syntax and grammar* (complexity, correctness), *vocabulary* and *language mechanics* (orthography, punctuation)

Appendix E

Table E. 1
Model fit metrics and unique variance explained by the predictors for the nine multi-level regression models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Dependent variable	Communi-cative effect	Level of detail	Coherence	Cohesion	Complexity syntax and grammar	Correctness syntax and grammar	Vocabulary	Orthography	Punctuation
Model fit indices									
Marginal R^2	0.28	0.30	0.30	0.31	0.40	0.45	0.49	0.45	0.40
Conditional R^2	0.40	0.43	0.52	0.42	0.52	0.58	0.58	0.62	0.53
ICC_{adj}	0.17	0.19	0.31	0.16	0.20	0.24	0.18	0.31	0.23
Unique variance explained: Semi-partial R^2 and unique share of marginal R^2 in % (only significant results are displayed)									
Communicative effect									
Level of detail		0.02 (6.7%)							
Coherence	0.01 (3.6%)		0.01 (3.4%)						
Cohesion				0.04 (12.9%)		0.01 (2.2%)			
Complexity syntax & grammar							0.01 (2.1%)		
Correctness syntax & grammar									
Vocabulary		0.01 (3.4%)	0.01 (3.4%)	0.02 (6.4%)	0.04 (9.9%)	0.02 (4.5%)	0.06 (12.4%)	0.02 (4.5%)	0.01 (2.5%)
Orthography	0.01 (3.6%)	0.01 (3.4%)		0.02 (6.4%)	0.03 (7.4%)	0.04 (9.0%)	0.04 (8.2%)	0.08 (18.0%)	0.01 (2.5%)
Punctuation			0.01 (3.4%)		0.01 (2.5%)	0.01 (2.2%)		0.01 (2.2%)	0.16 (40.3%)

Note. Marginal R^2: proportion of variance explained by fixed effects only. Conditional R^2: proportion of variance explained by both fixed and random effects. ICC_{adj} (Adjusted Intraclass Correlation Coefficient): proportion of total variance that is attributable to Level-2 differences after controlling for the fixed effects in the model. Semi-partial R^2: Proportion of variance uniquely explained by an individual predictor, after removing variance in that predictor that is shared with other predictors. Unique share of marginal R^2: proportion of the model's total explained variance (marginal R^2) that can be attributed solely to the individual predictor, excluding any variance that the predictor shares with other variables in the model.

Data Availability

The authors do not have permission to share data.

References

Augst, G., Disselhoff, K., Henrich, A., Pohl, T., & Völzing, P.-L. (2007). *Text – Sorten – Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter [Text – Genre – Competence. A true longitudinal study on the development of text competence at primary school age]*. Peter Lang.

Ayoobiyan, H., & Ahmadi, A. (2023). Detecting halo effects across rubric criteria in L2 writing assessment: A many-facet Rasch analysis. *Applied Research on English Language, 12*(1), 159–176. https://doi.org/10.22108/are.2022.132503.1848

Brinker, K., Cölfen, H., & Pappert, S. (2018). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden [Linguistic text analysis: An introduction to basic concepts and methods]* (8th ed.). Erich Schmidt Verlag.

Brock, R. (2015). *Handbuch Writing (2014 und 2015). Kommentierte Schreibperformanzen 2014 und 2015: Gesamtbeurteilungen. Richtlinien zur Bewertung von Schreibleistungen (6. und 7. Schulstufe)* [Writing handbook (2014 and 2015). Annotated writing performances 2014 and 2015: Overall assessments. Guidelines for the assessment of writing performances (6th and 7th grade)]. Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE). ⟨https://www.bifie.at/node/3148⟩.

Deutschschweizer Erziehungsdirektoren-Konferenz. (2016). *Lehrplan 21 – von der D-EDK Plenarversammlung am 31.10.2014 zur Einführung in den Kantonen freigegebene Vorlage*. Bereinigte Fassung vom 29.02.2016 [Curriculum 21 - template approved by the D-EDK plenary assembly on 31.10.2014 for introduction in the cantons. Revised version from 29.02.2016]. Deutschschweizer Erziehungsdirektoren-Konferenz. ⟨https://v-fe.lehrplan.ch/downloads.php⟩.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Griva, E., & Chostelidou, D. (2013). Writing skills and strategies of bilingual immigrant students learning Greek as a second language and English as a foreign language. *Reading & Writing, 4*(1), 1–9. https://doi.org/10.4102/rw.v4i1.31

Griva, E., Tsakiridou, H., & Nihoritou, I. (2009). A study of FL composing process and writing strategies employed by young learners. In M. Nikolov (Ed.), *Early learning of modern foreign languages. Processes and outcomes* (pp. 132–148). Multilingual Matters. https://doi.org/10.21832/9781847691477-012.

Grotjahn, R., & Kleppin, K. (2017). Typen und Funktionen der Evaluation von Schreibkompetenzen [Types and functions of writing assessment]. In B. Akukwe, R. Grotjahn, & S. Schipolowski (Eds.), *Schreibkompetenzen in der Fremdsprache. Aufgabengestaltung, kriterien-orientierte Bewertung und Feedback* (pp. 29–40). Narr Francke Attempto.

Hallet, W. (2016). *Genres im fremdsprachlichen und bilingualen Unterricht. Formen und Muster der sprachlichen Interaktion* [Genres in foreign language and bilingual teaching. Forms and patterns of linguistic interaction]. Klett, Kallmeyer.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(4), 228–250. https://doi.org/10.1016/j.asw.2012.06.003

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs.* Part one - general findings. ⟨https://www.ealta.eu.org/resources.htm⟩.

Hasselgreen, A., & Caudwell, G. (2016). Equinox. *Assessing the Language of Young learners*. https://doi.org/10.3138/9781781794760

Hasselgreen, A., Kaledaite, V., Maldonado Martín, N., & Pizorn, K. (2012). *Assessment of young learner literacy linked to the Common European Framework of Reference for Languages*. Council of Europe Publishing. ⟨http://www.ecml.at/tabid/277/PublicationID/63/Default.aspx⟩.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*(2), 91–98. https://doi.org/10.1016/0742-051X(87)90010-2

Hemmerich, W. (2016). *StatistikGuru: Rechner zur Adjustierung des α-Niveaus* [StatistikGuru: Calculator for adjusting the α-level]. ⟨https://statistikguru.de/rechner/adjustierung-des-alphaniveaus.html⟩.

Henik, A., & Tzelgov, J. (1985). Control of halo error: A multiple regression approach. *Journal of Applied Psychology, 70*(3), 577–580. https://doi.org/10.1037/0021-9010.70.3.577

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. ⟨http://www.jstor.org/stable/4615733⟩.

Hyland, K. (2004). *Genre and second language writing*. University of Michigan.

Jacobs, R., & Kozlowski, S., W. J. (1985). A closer look at halo error in performance ratings. *The Academy of Management Journal, 28*(1), 201–212. https://doi.org/10.2307/256068

Jansen, T., Vögelin, C., Machts, N., Keller, S., & Möller, J. (2021). Don't just judge the spelling! The influence of spelling on assessing second-language student essays. *Frontline Learning Research, 9*(1), 44–65. https://doi.org/10.14786/flr.v9i1.541

Konrad, E., Holzknecht, F., Schwarz, V., & Spöttl, C. (2018). *Assessing writing at lower levels: Research findings, task development locally and internationally, and the opportunities presented by the extended CEFR descriptors*. British Council. ⟨https://www.britishcouncil.org/exam/aptis/research/publications/assessing-writing-lower-levels⟩.

Lai, E.R., Wolfe, E.W., & Vickers, D.H. (2012). *Halo effects and analytic scoring. A summary of two empirical studies.* Research report. Pearson.

Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2015). Differentiation of illusory and true halo in writing scores. *Educational and Psychological Measurement, 75*(1), 102–125. https://doi.org/10.1177/0013164414530990

Lewis, M. A. (2005). Towards a lexical view of language – a challenge for teachers. *Babylonia, 3*(05), 7–10.

LimeSurvey GmbH. (2021). *LimeSurvey: An Open Source survey tool*. LimeSurvey GmbH. ⟨https://www.limesurvey.org⟩.

Linacre, J.M. (2025). *A user's guide to FACETS: Rasch-model computer programs* [Software manual]. ⟨https://www.winsteps.com/manuals.htm⟩.

Lindgren, E., & Stevenson, M. (2013). Interactional resources in the letters of young writers in Swedish and English. *Journal of Second Language Writing, 22*(4), 390–405. https://doi.org/10.1016/j.jslw.2013.09.001

Lohmann, J. F., Lötscher, F., Keller, S. D., Fleckenstein, J., Jansen, T., & Möller, J. (2025). Testing teacher judgment effects comprehensively: Accuracy, halo, frame of reference, strategy, and personality effects in the assessment of student essays. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000969

Lohmann, J. F., Machts, N., Möller, J., & Zitzmann, S. (2025). A more comprehensive, more reliable multilevel approach for assessing and modeling teacher judgment accuracy using latent variables. *Educational Psychology Review, 37*(2), 53. https://doi.org/10.1007/s10648-025-10029-z

Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., & Reble, R. (2022). Judgment accuracy of German student texts: Do teacher experience and content knowledge matter? *Teaching and Teacher Education, 119*, 1–8. https://doi.org/10.1016/j.tate.2022.103879

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19*(2), 139–158. https://doi.org/10.1080/10573560308222

Patekar, J. (2021). A look into the practices and challenges of assessing young EFL learners' writing in Croatia. *Language Testing, 38*(3), 456–479. https://doi.org/10.1177/0265532221990657

Péry-Woodley, M.-P. (1991). Writing in L1 and L2: analysing and evaluating learners' texts. *Language Teaching, 24*(2), 69–83. https://doi.org/10.1017/S0261444800006170

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Sun, T., Wang, C., Lambert, R. G., & Liu, L. (2021). Relationship between second language English writing self-efficacy and achievement: A meta-regression analysis. *Journal of Second Language Writing, 53*, Article 100817. https://doi.org/10.1016/j.jslw.2021.100817

Szabo, T. (2016a). *Collated representative samples of descriptors of language competences developed for young learners aged 7–10 years*. Council of Europe. ⟨http://www.coe.int/en/web/portfolio/overview-of-cefr-related-scales⟩.

Szabo, T. (2016b). *Collated representative samples of descriptors of language competences developed for young learners aged 11–15 years*. Council of Europe. ⟨http://www.coe.int/en/web/portfolio/overview-of-cefr-related-scales⟩.

Tono, Y., & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages. *International Journal of Corpus Linguistics, 19*(2), 163–177. https://doi.org/10.1075/ijcl.19.2.01ton

Trüb, R., Möller, J., Lohmann, J., Jansen, T., & Keller, S. D. (2025). Judgment accuracy in primary school EFL writing assessment: Do text characteristics matter? *Assessing Writing, 66*, 100957. https://doi.org/10.1016/j.asw.2025.100957

Trüb, R. (2022). An empirical study of EFL writing at primary school. Narr Francke Attempto. doi:10.24053/9783823395430.

UNESCO Institute for Statistics. (2012). *International standard classification of education. ISCED 2011*. ⟨http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf⟩.

Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal, 49*(6), 631–647. https://doi.org/10.1080/09571736.2018.1522662

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

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2020). Organisational quality of ESL argumentative essays and its influence on pre-service teachers' judgments. *Cogent Education, 7*(1), 1–21. https://doi.org/10.1080/2331186X.2020.1760188

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402. https://doi.org/10.1080/15434303.2014.960046

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357–378. https://doi.org/10.1037/a0022936

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*(1), 21–39. https://doi.org/10.1007/s10212-014-0225-6

Ruth Trüb is ad interim professor at the Institute for Primary Education at the University of Applied Sciences and Arts Northwestern Switzerland. Her research interests are EFL writing, teachers' assessment competence, English language teaching methodology and cross-linguistic transfer.

Julian Lohmann is a postdoctoral researcher at the Leibniz Institute for Science and Mathematics Education in Kiel in the department for Educational Measurement and Data Science.

Jens Möller is professor of psychology at the Institute for Psychology of Learning and Instruction at Kiel University. His research interests are diagnostic competence, bilingual learning and motivation.

Stefan D. Keller is professor and head of the research department for Subject-Specific Education at the Zurich University of Teacher Education. He specialises in English language teaching research.