Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol. 9 No. 1 (2021) 44 - 65
ISSN 2295-3159

Don't Just Judge the Spelling! The Influence of Spelling on Assessing Second-Language Student Essays

Thorben Jansen[1], Cristina Vögelin[2], Nils Machts[1], Stefan Keller[2] & Jens Möller[1]

[1] Institute of Psychology of Learning and Instruction, Kiel University, Germany
[2] Institute for Educational Sciences, University of Basel, Switzerland

Article received 27 September 2019 / Article revised 17 December 2020 / Accepted 21 December / Available online 11 February 2021

Abstract

When judging subject-specific aspects of students' texts, teachers should assess various characteristics, e.g., spelling and content, independently of one another since these characteristics are indicators of different skills. Independent judgments enable teachers to adapt their classroom instruction according to students' skills. It is still unclear how well teachers meet this challenge and which intervention could be helpful to them. In Study 1, N = 51 pre-service teachers assessed four authentic English as a Second Language (ESL) essays with different overall text qualities and different qualities of spelling using holistic and analytic rating scales. Results showed a negative influence of the experimentally manipulated spelling errors on the judgment of almost all textual characteristics. In Study 2, an experimental prompt was used to reduce this judgment error. Participants who were made aware of the judgment error caused by spelling errors formed their judgments in a less biased way, indicating a reduction of bias. The determinants of the observed effects and their practical implications are discussed.

Keywords: teacher judgments; second language writing; halo effect; spelling, writing assessment

Corresponding author: Thorben Jansen; Institute for Psychology of Learning and Instruction (IPL) University of Kiel Olshausenstraße 75 24118 Kiel Germany. Email: tjansen@ipl.uni-kiel.de *DOI: https://doi.org/10.14786/flr.v9i1.541*

1. Introduction

For the last two decades, second language education has received increasing attention in Europe. European citizens are expected to be proficient in two second-languages in addition to their first language when they complete secondary education (European Commission, 2008). Thus, there is a need for high-quality second language instruction in schools and close monitoring of its effectiveness. An important part of second language education is writing instruction, as writing is a key competence for higher education in the globalized world in general and tertiary education in particular (Keller, 2013). The importance of writing, in turn, requires competent teachers who have a good grasp of text quality, know how to give helpful feedback, and can assign fair and objective judgments based on transparent criteria. Accurate teacher judgments of students' achievement are, therefore, an important aspect of effective instruction (Elliott, Lee, & Tollefson, 2001), especially when teachers align their lessons with the perceived competencies of their students (Brookhart, 2011, 2013; Herppich et al., 2017; Ruiz-Primo & Furtak, 2007). Also, accurate teacher judgments serve as the primary source of information for students to evaluate their performances and affect their self-perception (Zimmermann, Möller, & Köller, 2018).

In the case of assessing students' second-language essays, there is scant research on teacher judgments. Previous research on teacher judgments neglected subject- and genre-specific achievements and focussed on teachers' ability to judge students' competencies in general (Südkamp, Kaiser, & Möller, 2012). There are aspects specific to second language writing assessment which might distort teacher judgments (Vögelin, Jansen, Keller, Machts, & Möller, 2019). For example, teachers need to distinguish between several independent characteristics, such as spelling, content, or organization, when assessing students' essays (Bae & Bachman, 2010; Cooksey, Freebody, & Wyatt-Smith, 2007; Flower & Hayes, 1981; Hyland, 2008). Empirical studies have found that textual characteristics, such as the number of spelling errors, influence teacher judgments of other independent characteristics, which is defined as a halo effect (Murphy & Reynolds, 1988; Saal, Downey, & Lahey, 1980). The term halo effect describes the impact of one characteristic on the judgment of an independent characteristic. A halo effect occurs, for example, when the number of spelling errors affects content-related judgments of a text. Halo effects are particularly likely in the assessment of English as a Second Language (ESL) student texts since these texts typically contain more spelling errors than first-language student texts (Flor, Futagi, Lopez, & Mulholland, 2015), and because teachers focus more on those errors (Cumming, Kantor, & Powers, 2002). When writing in a second language, students face larger challenges at the formal level (spelling, grammar, etc.) in addition to the content-related level (content, organization, etc.). As a result, the difference between content-related and formal criteria of second-language texts is typically greater than in first-language texts (Hamp-Lyons, 1991; Weigle, 2002), and the correlation between judgments of these levels is typically lower in second-language than in first-language texts (Bae & Bachman, 2010; Wind, Stager, & Patil, 2017). However, no studies investigated those halo effects on the judgment of ESL student texts resulting from spelling. Such halo effects could have serious consequences for learners as they could lead to misjudgments of other writing aspects, such as content, organization, argumentation, or structure of an essay, and thus lead to students losing motivation (Urhahne, 2015; Zhou & Urhahne, 2013) and confidence (Artelt, 2016). Additionally, they could be a major problem in language instruction in school as students with a deficit or strength in formal language were in danger of being under- or overestimated by their teachers Additionally, they could be a major problem in language instruction in school as students with a weakness or strength at the formal level would be in danger of being under- or overestimated by their teachers. Students might receive wrong grades or teachers might not plan subsequent lessons adequately when aligning their teaching to students' perceived competencies. In the worst case, students could be assigned to unsuitable educational tracks based on biased scoring (Schrader, 2013). Furthermore, concerning many written high-stakes exams in foreign education in secondary school and standardized second-language tests, halo effects from spelling could be crucial for certain students erroneously not being admitted to tertiary education.

Recent reviews and meta-analyses on teachers' judgment accuracy have stressed the urgent need for intervention studies on improving the quality of teachers' judgment (Kaufmann, 2020; Urhahne & Wijnia, 2021). Facilitating diagnostic competencies is a relatively new field with only nine empirical studies, none of which contain concrete interventions or suggesting methods to foster the quality of teachers' judgments on students' written performances (Chernikova et al., 2019). To strengthen the literature, we used a two-step approach to detect a judgment bias in the first study and reduce it in the second. The two studies presented in this paper represent each of the steps. In Study 1, we examined whether the number of spelling errors in ESL argumentative essays influences pre-service teachers' assessments of other textual characteristics using genre-specific analytic rating scales. In Study 2, we adapted an intervention from the research of professional rater training to the context of teacher education to investigate whether distortions in judgment can be reduced using a prompt alerting teachers' to possible halo effects.

2. The Influence of Spelling Errors on the Assessment of Students' Texts

Teacher assessment of students' achievement is moderated by (i.e., varies as a function of) a number of variables. The heuristic model of judgment accuracy by Südkamp et al. (2012) systematizes these moderators of teacher judgment accuracy, which is defined as the correspondence between teacher judgments of students' achievement and students' actual achievement. The model systematizes the moderators into four factors: it differentiates teacher characteristics (e.g., teaching experience, specialist knowledge, and pedagogical knowledge), test characteristics (e.g., relevant characteristics of a test, genre of a test, reliability of a test), student characteristics (e.g., performance, gender, motivation, age), and characteristics of the judgment to be made (e.g., the specificity of the domain to be judged, number of levels on a rating scale). When it comes to assessing writing, a special judgment characteristic is that teachers should use multiple pre-set scoring criteria and make the criteria available to the students (Huot, 1996). Hence, it is necessary to define which text characteristics should and should not influence raters' judgments a rating scale. Raters considering text characteristics in the way that is defined on the rating scale contributes to a correct judgment, whereas considering characteristics contrary to a rating scale – or confusing different aspects – contributes to a distorted judgment. For example, the number of spelling errors should influence the judgment of general text quality and spelling of a text; however, it should not influence the judgments of the content or organization (Parr & Timperley, 2010).

The heuristic model of judgment accuracy by Südkamp et al. (2012) describes which factors moderate teacher judgment accuracy. However, because the described model is a heuristic model, it does not describe the judgments' estimation process. It is useful to investigate the estimation process of judgments to investigate which factors influence judgment accuracy. Dual-process models of social information processing describe the estimation and the information processing that underlies it (Herppich et al., 2017; Karst, Dotzel, & Dickhäuser, 2018), such as the *Continuum Model of Impression Formation Processes* (Fiske & Neuberg, 1990). This model describes individuals' information processing as a continuum between a heuristic strategy and a controlled strategy to process information. The heuristic or shortcut strategy is assumed to be relatively automatic and to require little cognitive effort. For example, Fiske and Neuberg (1990) showed that individuals tend to use more heuristic information processing when they interpret target attributes to fit well into a category. Using this strategy leads to a simplification of the information in order to handle the complexity of the judgment more easily, thus increasing the possibility of halo effects. Using the controlled strategy, the individual collects more information and uses algebraic rules to integrate the information into a judgment. This strategy should minimize halo effects and will be used if individuals attend closely to target attributes.

Empirical studies have shown that teachers collect information about text characteristics in student texts and integrate them correctly into their judgment. Student texts with few spelling errors were assessed more positively (Birkel & Birkel, 2002) compared to texts with numerous spelling errors.

Teachers also assess texts' general quality more positively if they contained fewer errors, regardless of whether participants read only one (Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010) or 24 texts (Barkaoui, 2010).

Further, empirical studies have shown that textual characteristics influencing the judgment on scales that should not be influenced by the respective text characteristics indicate heuristic information processing. For example, teachers who were supposed to judge the content only assessed texts with several spelling errors as more negative concerning content than texts with fewer spelling errors (Marshall, 1967; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966). Scannell and Marshall (1966) and Marshall (1967) asked students and teachers to judge the overall quality of a history essay solely based on its contents. The authors prepared several versions of the essay, which only differed in spelling, punctuation, and grammatical errors. Texts with more than three spelling errors per 100 words were assessed more negatively with regard to content than texts with three or fewer spelling errors. Similarly, Rezaei and Lovorn (2010) found that the number of spelling, grammatical, and punctuation errors influenced the judgment of the content, although teachers had been explicitly instructed to assess content only. The influence of spelling errors was also seen in the judgment of ESL student writing. Raforth and Rubin (1984) and Sweedler-Brown (1993) prepared different versions of one or six second-language essays by students that differed only in the number of spelling errors. The participants were supposed to assess content, organization, structure, and grammar in addition to the overall quality. In both studies, the number of spelling errors negatively affected the judgment of all assessment dimensions.

Thus, several studies have shown that the number of spelling errors influences the assessment of textual characteristics, even if they should not. However, no study exists on the assessment of ESL argumentative essays in an upper-secondary school context. The results of these studies are only applicable to a limited extent since the studies differ from an ESL school context concerning teachers' characteristics, students' characteristics, and the judgment to be made, which could influence the assessment (Südkamp et al., 2012). In many studies, the texts did not originate from pupils but from university students (Barkaoui, 2010; Freedman, 1979; Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010; Sweedler-Brown, 1993) and professional raters but teachers assessed them (Freedman, 1979; Rafoth & Rubin, 1984; Sweedler-Brown, 1993; Wolfe, Song, & Jiao, 2016). Moreover, unlike in a school context, participants lacked the opportunity to compare texts because they only read a single text (Marshall, 1967; Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966).

This paper presents two studies focusing on fostering teachers' diagnostic competencies by reducing halo-effects caused by spelling errors on other aspects in the assessment of advanced argumentative L2 essays from upper-secondary school. We examined two sequential research questions with two sequential studies using the same material. Firstly, do halo effects distort pre-service teachers' ratings of students' texts? Secondly, can a prompt reduce this halo-effect? Study 1 addressed the phenomenon in an experimentally controlled within-subject design, examining how different numbers of spelling errors affected the judgment on rating scales that should and should not be influenced by spelling. Study 2 examined whether a prompt could reduce the erroneous influence of spelling errors. As we show in detail below, we divided the participants in Study 2 into a control and an intervention group. Participants in both groups were presented the same material as in Study 1. Participants in the intervention group additionally received a prompt that alerted them to possible halo-effects arising from the number of spelling errors.

3. Study 1

This study aimed to investigate whether and to what extent spelling influences the assessment of different textual characteristics. For this, participants assessed four ESL argumentative essays of higher or lower overall quality in an experimental study. The research team prepared two versions of each text that differed only in the number of spelling errors. Pre-service teachers were then asked to evaluate these texts with a holistic scale for the texts' overall assessment and seven analytic scales with detailed descriptors of the different levels. We tested the following hypotheses:

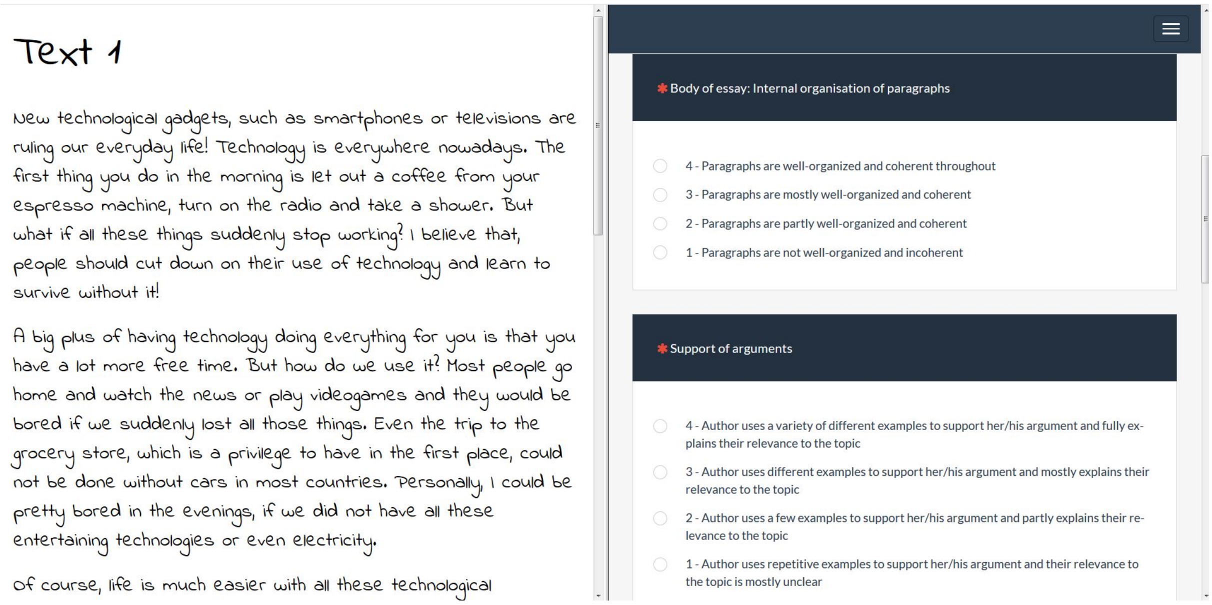
Hypothesis 1: Texts of low overall quality are assessed more negatively on the holistic and all analytic scales than texts of higher quality.

Hypothesis 2: Texts with more spelling errors are assessed more negatively than texts with few spelling errors on the scale that should be influenced by spelling.

Hypothesis 3: Texts with few spelling errors are assessed more positively than texts with many spelling errors on the scales that should **not** be influenced by spelling (halo effect).

3.1. Method

Participants assessed student texts in a digital instrument called the *Student Inventory ASSET* (see figure 1; Vögelin, Jansen, Keller, & Möller, 2018). This instrument was developed on the basis of earlier work by Kaiser, Möller, Helm, and Kunter (2015). In this computer-based tool, participants read student texts on the left-hand side and see the rating scales displayed on the screen's right-hand side. Each participant assessed four English student texts (two with high and two with low overall text quality as well as two with few spelling errors and two with many spelling errors). Texts were presented in a randomized sequence. Participants first assessed the texts' general quality on a holistic scale before they assessed seven individual characteristics on analytic scales.



Text 1

New technological gadgets, such as smartphones or televisions are ruling our everyday life! Technology is everywhere nowadays. The first thing you do in the morning is let out a coffee from your espresso machine, turn on the radio and take a shower. But what if all these things suddenly stop working? I believe that, people should cut down on their use of technology and learn to survive without it!

A big plus of having technology doing everything for you is that you have a lot more free time. But how do we use it? Most people go home and watch the news or play videogames and they would be bored if we suddenly lost all those things. Even the trip to the grocery store, which is a privilege to have in the first place, could not be done without cars in most countries. Personally, I could be pretty bored in the evenings, if we did not have all these entertaining technologies or even electricity.

Of course, life is much easier with all these technological

* Body of essay: Internal organisation of paragraphs

○ 4 - Paragraphs are well-organized and coherent throughout

○ 3 - Paragraphs are mostly well-organized and coherent

○ 2 - Paragraphs are partly well-organized and coherent

○ 1 - Paragraphs are not well-organized and incoherent

* Support of arguments

○ 4 - Author uses a variety of different examples to support her/his argument and fully explains their relevance to the topic

○ 3 - Author uses different examples to support her/his argument and mostly explains their relevance to the topic

○ 2 - Author uses a few examples to support her/his argument and partly explains their relevance to the topic

○ 1 - Author uses repetitive examples to support her/his argument and their relevance to the topic is mostly unclear

Figure 1. Assessment in the Student Inventory ASSET.

3.2. Sample

N = 51 pre-service teachers participated in this study. The required sample size for analyzing the effects within the subjects was calculated with G*Power (Faul, Erdfelder, Lang & Buchner, 2007). Based on other findings with the *Student Inventory ASSET* (Jansen, Vögelin, Machts, Keller, & Möller, 2019; Kaiser et al., 2015; Vögelin et al., 2019), a moderate effect size of *d* = 0.60 was expected so that, at a power of *B* = .90, the required sample size was *N* = 33. The samples consisted of pre-service English teachers who were recruited from master seminars at universities Kiel and Basel. The average age of the participants was *M* = 29.41 (*SD* = 7.25) years. 62.7% of the subjects were female.

3.3. Variables

We used a 2x2 experimental design with two independent variables that were varied within the subjects: overall quality (low vs. high) and a number of spelling errors (“few errors” vs. “many errors”). The dependent variables were the holistic and analytic assessments.

3.3.1. Text quality

Text quality varied in two levels: “high” and “low”. We used expert ratings to operationalize text quality, which is a common approach in writing research (Meadows & Billington, 2010; Royal-Dawson & Baird, 2009; Scanell & Marshall, 1966). Experts from the School of Teacher Education (location anonymized) evaluated 15 student texts together with the ASSET research team using the NAEP rating scale (Driscoll, Avallone, Orr, & Crovo, 2010). As a result, we chose two “stronger” and two “weaker” texts from the sample, each of roughly equivalent overall quality and adjusted them to the same text length. The weaker texts exhibited levels of work that received a failing grade, i.e., considered insufficient by all experts regarding the unit’s learning goals. The stronger texts that were chosen showed levels of work that were considered to have surpassed the learning goals and received good to excellent ratings by experts.

The texts came from students who had been learning English as a second language for five years and had been instructed for four weeks on the form, structure, and content of argumentative essays. At the end of the teaching unit, the students wrote an argumentative essay in 90 minutes answering the following writing prompt: “Do you agree or disagree with the following statement? As humans are becoming more dependent on technology, they are gradually losing their independence.” All texts were of similar length (between 450 and 465 words) since text length significantly influences text assessment (Wolfe et al., 2016).

3.3.2. Number of spelling errors

This study included both spelling and punctuation errors in our manipulation since relevant rubrics often list spelling and punctuation issues in the same category (usually “language mechanics”). We varied the variable within the texts similar to Raforth and Rubin (1984) and Sweedler-Brown (1993) in two steps (“few errors” vs. “many errors”). In the text version with few errors, spelling errors were reduced to one error per 100 words, whereas in the text version with many errors, spelling errors were raised to seven errors per 100 words. The numbers (frequency) of spelling errors were derived from the mean number of spelling errors in the high and low-quality texts of corpus overall. Thus, we aimed to ensure that the manipulation occurred within the range of the students’ language competencies. Table 1 displays the types of spelling and punctuation errors we integrated into text variations with a low quality of spelling.

Table 1

Types of spelling and punctuation errors

Spelling errors	**Examples**
Phonemic orthography	*articel, *chrismas
Wrong/no doubling of letters	*conected, *developpments
Wrong use of open, closed, and hyphenated compounds	*mobilephone, *world wide
Wrong capitalization	now *Diseases would spread faster
Wrong use of single letters	*midication, *correkt
Punctuation errors	**Examples**
Omission of full stop	*To conclude, technology makes you able to do many things The technology should give you a better life.
Wrong insertion of commas	*You feel dependent on technology, if you do not have any alternatives.

3.3.3. Text assessment

Participants assessed each text on the six-level holistic scale of the National Assessment of Educational Progress (Driscoll et al., 2010). This study further employed genre-specific analytic rating scales (for an overview, see Jansen, Vögelin, Machts, Keller, Köller, & Möller, 2021). These seven scales were based on the 6 + 1 trait model (Culham, 2003) as well as the Test in English for Educational Purposes (TEEP) (Weir, 1988), and we adjusted them to address genre-specific characteristics of argumentative essays (Hyland, 1990; Zemach & Stafford-Yilmaz, 2008). The seven scales were *frame of essay: introduction and conclusion, body of essay: internal organization of paragraphs, support of arguments, spelling and punctuation, grammar*, *vocabulary,* and *overall task completion*. Each dimension contained four levels with detailed descriptors, with higher levels indicating a positive assessment (see Appendix).

We used two two-factor multivariate variance analyses with repeated measurements (MANOVA) to analyze the data with subsequent post-hoc tests. In doing so, we conducted two separate MANOVA, one for the scales that should be influenced by spelling (*holistic scale*, *spelling and punctuation)* and one for the scales that should not be influenced by spelling (*frame of essay, body of essay, support of arguments, grammar, vocabulary,* and *overall task completion*). When an effect of spelling occurred on scales that should be influenced by the spelling, it showed that teachers were able to recognize the spelling errors. By contrast, the analyses were considered as halo-effects when an effect of spelling occurred on scales that should not be influenced by the spelling.

We also analyzed the data using nonparametric tests, which confirmed our results. Therefore, we only present the MANOVA results.

3.4. Results

Table 2 shows descriptive results of the assessments.

The analysis of the two scales that should be influenced by spelling (*holistic scale* and *spelling and punctuation*) showed significant multivariate main effects for spelling errors ($F(2, 49) = 66.03$, $p < .001$) and text quality ($F(2, 49) = 65.56$, $p < .001$), and no interaction between text quality and spelling errors ($F(2, 49) = 1.93$, *ns*) for judgments. Results of univariate post-hoc tests (see Table 2) showed effects for spelling and text quality on both scales. Low-quality texts were judged more negatively than high-quality texts, and texts with many spelling errors were judged more negatively than texts with few spelling errors. Our findings thus supported Hypotheses 1 and 2.

The analyses of the scales that should not be influenced by spelling (*frame of essay, body of essay, support of arguments, grammar*, *vocabulary,* and *overall task completion)* also showed significant multivariate main effects of text quality (*F*(6, 45) = 32.30, *p* < .001), spelling errors (*F*(6, 45) = 7.41, *p* < .001), and no interaction between text quality and spelling errors (*F*(6, 45) = 1.45, *ns)*. The results showed that texts of low quality were judged significantly more negatively on all scales than texts of high quality (see Table 2). This result further supported Hypothesis 1. The post-hoc analysis for the effect of spelling errors showed that texts with many spelling errors were judged more negatively than texts with few spelling errors on five out of six scales (the exception was *frame of essay*) that should not be influenced by spelling. These findings thus supported Hypothesis 3.

Table 2
Means, Standard Deviations, and Univariate Analyses of Variance in Study 1 for the Variables Spelling Errors ("Few Errors" vs. "Many Errors") and Text Quality ("High" vs. "Low")

Dependent Variables	Spelling Errors				Text Quality			
	few	many			high	low		
	M(SD)	*M(SD)*	*F (1, 50)*	*d*	*M(SD)*	*M(SD)*	*F (1,50)*	*d*
Scales that should be influenced by Spelling								
Holistic Scale	3.99 (0.76)	3.32 (0.78)	18.94***	0.87	4.40 (0.77)	2.91 (0.74)	100.68***	1.97
Spelling and Punctuation	3.18 (0.55)	1.78 (0.56)	131.24***	2.52	2.88 (0.53)	2.08 (0.43)	73.12***	1.66
Scales that should **NOT** be influenced by Spelling								
Frame of Essay	2.72 (0.62)	2.55 (0.56)	1.80	0.29	3.03 (0.55)	2.24 (0.54)	54.35***	1.45
Body of Essay	2.78 (0.57)	2.46 (0.55)	11.16**	0.57	3.05 (0.55)	2.20 (0.58)	73.77***	1.50
Support of Arguments	2.84 (0.50)	2.64 (0.50)	4.14*	0.32	3.13 (0.50)	2.35 (0.59)	44.14***	1.43
Grammar	2.82 (0.54)	2.24 (0.51)	30.58***	0.86	3.13 (0.48)	1.93 (0.50)	169.37***	2.45
Vocabulary	2.74 (0.47)	2.31 (0.53)	30.86***	1.10	3.16 (0.61)	1.89 (0.49)	155.82***	2.30
Overall Task Completion	2.73 (0.39)	2.44 (0.47)	9.32**	0.67	3.00 (0.51)	2.17 (0.36)	74.30***	1.88

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

3.5. Discussion

Study 1 showed that pre-service teachers correctly differentiated between two quality levels of spelling and overall text quality when evaluating learners' writing competencies. This result is in line with previous studies showing similar results: teachers collect information about text characteristics in student texts and integrate them correctly into their judgment (Birkel & Birkel, 2002; Barkaoui, 2010; Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010). More importantly, it demonstrated that spelling errors had a considerable influence on teacher judgments when they assessed text characteristics, which should be evaluated separately from spelling errors. All text characteristics, besides spelling and punctuation, were identical in both versions, yet they were assessed more negatively when texts included more spelling errors. Pre-service teachers' judgments regarding the characteristics *body of essay, support of arguments, grammar*, *vocabulary,* and *overall task completion* were all subject to halo effects. According to the *Continuum Model* (Fiske & Neuberg, 1990), these halo effects could indicate the use of a heuristic information processing strategy. We assume that to assess these characteristics, the pre-service teachers needed to focus on all areas of the text simultaneously, which made it hard to pay attention to the individual aspects of assessment and thus more difficult to distinguish between the analytic criteria. Only the characteristic *frame of essay* (i.e., whether a text had a clear introduction and conclusion) was not influenced by the number of spelling errors, possibly indicating a stronger use of the controlled information processing strategy. The category *frame of essay* may be less prone to halo effects since its assessment is limited to the introduction and conclusion of the text, which made it easier to attend focus on the specific descriptors while excluding distractions of spelling. This might explain why pre-service teachers employed more controlled information processing when assessing this particular text characteristic.

One could debate whether it is theoretically possible for teachers—or raters—to disentangle different rating criteria completely (Sadler, 2009). While rating criteria should be non-overlapping, there is ample research showing that empirical scores for rating criteria intercorrelate (Huot, 1996). In terms of formative feedback and judgment validity, this overlap has limits, however. When the quality of spelling negatively influences the assessment of structure or argumentation in an essay, this indicates a halo-effect and a distortion of judgment, which is in line with previous research (Marshall, 1967; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966). Telling a student to improve the overall quality of a text by correcting the spelling is sound advice. Telling a student that her essay uses poor argumentation when the problem is poor spelling, by contrast, is unsuitable or misleading feedback. Therefore, it is a key task for teacher education to alert teachers to such distortion effects and show them ways of avoiding or reducing them. According to the *Continuum Model* (Fiske & Neuberg, 1990), the use of the heuristic strategy can be reduced if evaluators spend more attention on target attributes of the assessment, possibly preventing halo effects. This could be encouraged by a prompt instructing evaluators to assess every analytic scale as a different, independent category. Previous studies show that additional attention to the assessment's target attributes increases the reliability (Lovorn & Rezaei, 2011) and accuracy (Chamberlain & Taylor, 2011; Dempsey, PytlikZillig, & Bruning, 2009; Meadows & Billington, 2010) of the assessment based on analytic scales. A recent meta-analysis on fostering diagnostic competence of pre-service teachers (Chernikova et al., 2019) identified large positive effects for the prompts on pre-service teachers' judgments. However, the meta-analysis contained only nine studies, including no study fostering the assessment of students' written performances. To fill this research gap, Study 2 examines whether a prompt could reduce the halo-effect of spelling.

4. Study 2

Study 2 tested an intervention to reduce the halo effect found in Study 1. The participants in Study 2 were randomly split into a control group, which replicated Study 1, and an experimental group. Before the assessment, the participants in the experimental group were given a verbal prompt that instructed them to pay attention to the possible influence of spelling errors on their assessment of other text characteristics. This study tested the following hypotheses:

Hypothesis 1: Texts of low overall quality are assessed more negatively on the holistic and all analytic scales than texts of higher quality.

Hypothesis 2: Texts with more spelling errors are assessed more negatively than texts with few spelling errors on the scale that should be influenced by spelling.

Hypothesis 3: Texts with few spelling errors are assessed more positively than texts with many spelling errors on the scales that should **not** be influenced by spelling (halo effect).

Hypothesis 4: The prompt reduces the halo effect of spelling.

4.1. Method

4.2. Sample

$N = 66$ pre-service teachers participated in Study 2. The required sample size for analyzing the interaction between effects within and between the subjects was calculated with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). The halo effect size ($d = 0.63$) averaged across all scales found in Study 1 was expected so that, at a power of $B = .90$, a sample size of $N = 54$ was required. The samples consisted of pre-service teachers of English, who were recruited from seminars at universities (locations anonymized). The average age of the participants was $M = 24.31$ ($SD = 4.90$) years, and 66% were female. The sample group was randomly split into two groups: the “prompt group” ($N = 30$) and the “control group” ($N = 36$).

4.3. Variables

4.3.1. Text quality

We employed the same four texts of two quality levels that were used in Study 1.

4.3.2. Number of spelling errors

As in Study 1, we varied the number of spelling errors at two levels: “few errors” (1 error per 100 words) and “many errors” (7 errors per 100 words).

4.3.3. Prompt

In the *Student Inventory ASSET*, participants of the treatment group were shown the following prompt: “Empirical studies have shown that teachers are influenced to a large degree by spelling errors while assessing text quality” before being asked to assess the student texts. The control group saw the prompt “Please judge the texts in a balanced and fair way”. The first prompt was thus specifically aimed at alerting participants to possible halo effects emanating from spelling, while the second prompt asked for fair and unbiased assessment only in a very general fashion.

4.3.4. Text assessment

As in Study 1, participants assessed each of the four texts on the six-level holistic scale of the National Assessment of Educational Progress (Driscoll et al., 2010) and on seven four-level genre-specific analytic scales (Culham, 2003; Weir, 1988).

4.4. Results

We used two three-factor multivariate variance analyses with repeated measurements to analyze the data with subsequent contrast tests. In doing so, we reported the results separately for the scales that should or should not be influenced by spelling. Table 3 shows the descriptive parameters of Study 2.

The analyses of the scales that should be influenced by spelling (*holistic scale* and *spelling and punctuation*) showed significant multivariate main effects for the number of spelling errors ($F(2, 63) = 146.40$, $p < .001$) and the overall quality of the texts ($F(2, 63) = 97.60$, $p < .001$), and no main effect of the prompt ($F(2, 63) = 1.13$, *ns*). The multivariate analysis further showed no interaction effects for the text quality and the number of spelling errors ($F(2, 63) = 0.24$, *ns*), the text quality and the prompt ($F(2, 63) = 0.15$, *ns*), the number of spelling errors and the prompt ($F(2, 63) = 2.46$, *ns*), and no three-way interaction between the factors of spelling errors, quality, and prompt ($F(2, 63) = 1.58$, *ns*). The results of the post-hoc analyses showed univariate effects for spelling and for text quality on both scales in the assumed direction (see Table 4). As in Study 1, the results supported Hypothesis 1 and 2.

Additionally, the analyses of the scales that should not be influenced by spelling (*frame of essay, body of essay, support of arguments, grammar*, *vocabulary,* and *overall task completion)* showed significant multivariate main effects for the number of spelling errors ($F(6, 59) = 28.43$, $p < .001$), the quality of the texts ($F(6, 59) = 48.43$, $p < .001$), and the prompt ($F(6, 59) = 1.32$, *ns*).

Most importantly, as expected, there was also a significant interaction between the number of spelling errors and the prompt ($F(6, 59) = 2.34$, $p < .05$). The multivariate analysis showed no interaction between the spelling errors and the quality ($F(6, 59) = 0.66$, *ns*), the quality and the prompt ($F(6, 59) = 0.64$, *ns*), or the number of spelling errors, the quality, and the prompt ($F(6, 59) = 1.42$, *ns*).

Univariate post-hoc tests were calculated for significant multivariate effects on the scales that should not be influenced by spelling (see Table 4). For text quality, results indicated that texts of low overall quality were assessed more negatively on all scales than texts of high quality. These findings supported Hypothesis 1. The effect sizes were, again, consistently high. Similarly, texts with many spelling errors were assessed more negatively on all scales than were texts with few spelling errors. The data in this study also supported the appearance of halo effects formulated in Hypothesis 3: Pre-service teachers' judgments on all scales were influenced by the number of spelling errors.

The test for the effect of the prompt (Hypothesis 4) was a one-sided, post-hoc analysis of the interaction between spelling and prompt. It showed significant results on the scales *support of arguments*, *vocabulary*, and *overall task completion*. The prompt reduced the halo effect of the number of spelling errors on these scales (see Table 3).

Table 3

Means and Standard Deviation in Study 2 for the Variables Spelling Errors ("Few Errors" vs. "Many Errors") and Text Quality ("High" vs. "Low"); Split for the Prompt Group and the Control Group

Dependent Variables	Spelling Errors Prompt Group		Spelling Errors–Control Group		Text Quality–Prompt Group		Text Quality–Control Group	
	Few	Many	Few	Many	High	Low	High	Low
Scales that should be influenced by Spelling								
Holistic Scale	4.17 (0.78)	3.63 (0.71)	4.68 (0.81)	3.12 (0.74)	4.25 (0.70)	3.21 (0.71)	4.58 (0.77)	2.88 (0.64)
Spelling and Punctuation	3.30 (0.48)	1.78 (0.57)	2.88 (0.54)	2.20 (0.41)	3.24 (0.54)	1.64 (0.46)	2.79 (0.45)	2.08 (0.51)
Scales that should **NOT** be influenced by Spelling								
Frame of Essay	2.82 (0.52)	2.62 (0.50)	3.18 (0.56)	2.25 (0.45)	2.93 (0.52)	2.43 (0.60)	3.13 (0.57)	2.24 (0.44)
Body of Essay	3.07 (0.43)	2.62 (0.52)	3.30 (0.45)	2.38 (0.54)	2.83 (0.52)	2.35 (0.50)	3.03 (0.58)	2.15 (0.53)
Support of Arguments	2.72 (0.43)	2.65 (0.51)	3.03 (0.47)	2.33 (0.48)	3.00 (0.48)	2.39 (0.55)	3.07 (0.47)	2.32 (0.59)
Grammar	3.00 (0.51)	2.28 (0.43)	3.17 (0.56)	2.12 (0.47)	2.93 (0.50)	2.08 (0.46)	3.15 (0.48)	1.86 (0.44)
Vocabulary	2.83 (0.53)	2.43 (0.45)	3.23 (0.49)	2.03 (0.56)	2.86 (0.44)	2.17 (0.51)	3.17 (0.52)	1.86 (0.49)
Overall Task Completion	2.80 (0.39)	2.55 (0.51)	3.13 (0.56)	2.22 (0.36)	2.90 (0.41)	2.38 (0.53)	3.14 (0.52)	2.14 (0.53)

Table 4

*Post-Hoc Analyses of the Multivariate Spelling, Text Quality, and Prompt*Spelling Effects.*

Dependent Variables	Spelling Errors		Text Quality		Prompt* Spelling Errors	
	F (1, 64)	*d*	*F* (1, 64)	*d*	*F* (1, 64)	*d*
Scales that should be influenced by Spelling						
Holistic Scale	42.14***	1.11	166.28***	2.27		
Spelling and Punctuation	288.88***	3.08	75.32***	1.29		
Scales that should **NOT** be influenced by Spelling						
Frame of Essay	13.76***	0.67	120.24***	1.81	2.53	0.40
Body of Essay	37.58***	0.93	101.17***	1.67	0.06	0.06
Support of Arguments	16.08***	0.71	67.65***	1.44	10.38**	0.80
Grammar	120.53***	1.66	235.72***	2.41	0.84	0.23
Vocabulary	52.62***	1.14	209.12***	2.45	3.81*	0.48
Overall Task Completion	27.14***	0.88	120.07***	1.96	3.46*	0.46

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

4.5. Discussion

Study 2 aimed to test whether a prompt was effective at reducing halo-effects of spelling on other independent aspects of student writing, thus improving the quality of their assessments. Participants saw a prompt that alerted them to possible distortion effects of spelling before the assessment. We first examined whether the effects of text quality and spelling errors, including halo-effect, occurred in the same way as in Study 1. Results showed that pre-service teachers considered text quality and spelling errors when assessing ESL argumentative essays, replicating results from Study 1: texts of low overall quality, and many spelling errors, were assessed more negatively with regard to these two characteristics than texts of high quality and with few spelling errors. Moreover, the results showed that the number of spelling errors influenced teachers' judgments on all scales that should not be influenced by spelling errors. Hence, a halo effect was seen on all assessment scales and across all participants. These findings are in line with Study 1. This replication is a particular strength of the study and an important part of good scientific practice, especially in the so-called "replication crisis" (Bakker, van Dijk, & Wicherts, 2012; Open Science Collaboration, 2015).

Study 2 also expanded Study 1 by implementing a prompt to reduce halo effects. The prompt had an effect on participants' assessment of the scales *support for arguments, vocabulary,* and *overall task completion*. However, it did not have a significant effect on the assessment of *frame of essay, body of essay*, and *grammar.* Interpreting this fact is one of the more challenging aspects of the study. We used the *Continuum Model of Impression Formation Processes* (Fiske & Neuberg, 1990) as a theoretical model, in which information processing is described as a continuum between a heuristic and a controlled strategy to process information. The model predicts that individuals will use more heuristic information processing when target attributes are interpreted to fit well into a category. One could argue that participants found it easier to judge *support for arguments, vocabulary,* and *overall task completion* because they had clear notions of the amount of support, number of topic-specific words, or conventions that were required for this type of text. This would mean that it was possible for them to apply the controlled strategy of information processing and still handle the complexity of the judgment on these scales. By contrast, one could argue *frame of essay, body of essay,* and *grammar* were more difficult to judge because they were less clearly defined or less familiar to this type of participant (pre-service teachers). Thus, these scales would have been too complex for using the controlled strategy, rendering the prompt less effective. This interpretation is in line with the finding that error-explaining prompts are not effective when teachers' cognitive load is too high (Heitzmann, Fischer, & Fischer, 2018) or for teachers with low professional knowledge (Chernikova et al., 2019). In both cases, teachers need more instructional guidance, which would be especially relevant for pre-service teachers such as the ones who participated in our study. However, more studies are needed to investigate what makes scales too complex to use controlled information processing, and what knowledge teachers might require to reduce the complexity of judgments.

5. General Discussion

The two studies aimed to examine two sequential research questions: Do halo-effects of spelling errors distort pre-service teachers' ratings of students' texts? Can a prompt reduce this halo-effect? Regarding the first research question, the results in Study 1 and 2 showed halo-effects of spelling errors distorting pre-service teachers' ratings of students' texts. This finding is in line with our hypothesis and the research on the influence of spelling errors on the assessment of students' texts (Marshall, 1967; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966). Further, our studies presented three novelties that complement existing research: First, the sample consisted of pre-service teachers rather than of professional raters. Halo-effects in professional rating settings are a well-known phenomenon, and various interventions are used to safeguard against their distorting consequences (see Myford & Wolfe, 2009). In contrast, no safeguards exist for pre-service teachers, and hence, this judgment error could persist until teachers start working in a school. Based on our findings, the strong need to examine possibilities of reducing pre-service teachers' halo-effects becomes evident. Second, in comparison to previous research (Marshall, 1967; Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966), teachers assessed texts with the same overall quality but different amounts of spelling errors in our studies. During the text assessment, the participants saw examples of high-quality texts with few and many spelling errors and thus were able to experience these as two related but ultimately distinct categories of quality. Nevertheless, the halo-effect occurred and extended to textual aspects clearly unrelated to spelling, such as the framing of an essay or the internal structure of paragraphs. Seeing that even high-quality texts can contain spelling errors does not seem to protect pre-service teachers against halo-effects. The third novelty is that the texts analysed in our study were written by upper-secondary school students and not by university students, which indicates that the halo-effects of spelling errors are a more widespread problem than previous research has suggested (Barkaoui, 2010; Freedman, 1979; Rafoth & Rubin, 1984; Rezaei & Lovorn, 2010; Sweedler-Brown, 1993). Following these three novelties, the conclusion for our first research question is the following: halo-effects are a problem in upper-secondary ESL writing education and there is a need for an intervention to reduce it. We addressed this need in our second study.

The results of the second study showed that a prompt could reduce the halo-effect of spelling. This result is in line with our hypotheses, the assumptions of the *Continuum Model of Impression Formation Processes* (Fiske & Neuberg, 1990) and the research of fostering diagnostic competencies of pre-service teachers. We have chosen the prompt because of the assumption, derived from the *Continuum Model*, that drawing attention to target attributes of the assessment can reduce halo-effects. We interpret the effect of the prompt as support for this assumption and the usefulness of the *Continuum Model* to describe judgment processes. The prompt's effect size was similar to the mean prompt's effect size in a meta-analysis on fostering teachers' diagnostic competencies (Chernikova et al., 2019). However, in studies summarized in meta-analyses, the prompts were usually part of larger interventions. The results of our study contradict those of the only other study investigating prompting effects solely on facilitating teachers' diagnostic competencies (Heitzmann et al., 2018), which showed that prompts alone tend to hamper the quality of teachers' diagnoses and are only effective in combination with adaptive feedback. Heitzmann et al. (2018) explained the hampering effects with the high cognitive load teachers experienced when processing the prompt. Hence, it could be more efficient and economical to use prompts that elicit little cognitive load, like in the actual study.

The research in the relatively new field of fostering pre-service teachers' diagnostic competencies opens a promising avenue for further research. It contains a simple but innovative suggestion for fostering the quality of teacher assessments of written students' performances, combining interdisciplinary research from second language learning research and psychology, and showing beneficial prompting effects without additional adaptive feedback (Heitzmann et al., 2018). 'The particular strength of our study is that it showed a way of fostering the assessment quality that could be integrated into teacher education economically. Further, our results can help teachers in a way that encourages them to ask themselves: Did I misjudge my students' who have spelling difficulties, and did I overlook their strengths because of their spelling errors? We concluded for both research questions that domain-specific judgment characteristics, like analytic writing assessments, could trigger judgment errors, which can and should be detected and reduced to raise second language instruction quality.

A limitation of our study is that we did not randomize the order of holistic and analytic scoring in the studies. We used the scoring order recommended by Singer and LeMahieu (2011) and did not want to include a scoring order leading to more distracting interferences among the different scales. We included both a holistic rating scale and an analytic rating scale since both scales are commonly employed in practice (Weigle, 2002), and our study aimed to investigate halo effects in relation to achievement-relevant and achievement-irrelevant characteristics.

The experimental setting used in the *Student Inventory ASSET* assures internal validity due to its strict variable control and randomized allocation of text qualities and spelling errors. However, one should keep in mind that real-life assessment situations differ considerably from this simplified experimental research design (Keller, 2013). Hence, caution should be exercised when applying these insights into real-life situations.

Another limitation is the text selection: Four texts on two quality levels and with two independent degrees of spelling errors were assessed in the studies. While this selection made it possible to reach a high internal validity of our study, it makes it difficult to generalize the results, and it remains unclear whether prompts are also effective in school settings. In school, teachers have a larger amount of texts available that cover a wide continuum of qualities and in which the quality levels and the number of spelling errors correlate with one another (Bae, Bentler, & Lee, 2016; Lai, Wolfe, & Vickers, 2015).

Furthermore, the sample in our studies consisted of pre-service ESL teachers only. The applicability of the results to experienced teachers is limited since experienced teachers may discriminate between assessment scales more accurately due to their greater knowledge and experience, and hence may be less influenced by halo effects. However, to this date, no study demonstrated the difference between pre-service and experienced teachers' judgments in assessing student texts (Meadows & Billington, 2010; Royal-Dawson & Baird, 2009). We decided to conduct our study with pre-service teachers to maximize the possible outreach of the prompt: if the prompt reduced halo-effects in teacher' education, the effects could be beneficial throughout teachers' professional career. To generalize our findings, we encourage other researchers to conduct similar experiments with experienced teachers.

For the scientific discussion of performance assessment, previous research and the outlined studies indicate that the influence of spelling errors should also be investigated in other subjects, such as science education. Further studies could systematically investigate other forms of support that help teachers reduce halo effects in the assessment.

In practice, results indicate that teachers judge the quality of student texts and thus the competencies of students in certain areas more erroneously when texts contain many spelling errors. These judgment errors can both systematically disadvantage students and interfere with competence-adjusted lesson planning. Thus, it is important to inform pre-service and experienced teachers of typical difficulties in writing assessment and support them in assessing students' written performances objectively. Hence, assessment scales and training programs such as those offered by the *Student Inventory ASSET* should be further improved and used in empirical studies of the complex psychological processes underlying text assessment. Further, they should be made available to both pre-service and experienced teachers as training instruments.

Key points

- Examining teachers' holistic and analytic writing assessment
- English as a Second-Language (ESL) essays from upper-secondary school
- Spelling errors triggered halo effects on analytic rating scales
- A prompt could reduce these halo effects
- The first study shows the judgment error; the second reduces it

Funding

This work was supported by the Swiss National Science Foundation [165483; 100019L162675] and the German research foundation (DFG) [Mo648/25-1]

References

Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing*, *27*(2), 213–234. https://doi.org/10.1177/0265532209349470

Bae, J., Bentler, P. M., & Lee, Y.-S. (2016). On the role of content in writing assessment. *Language Assessment Quarterly*, *13*(4), 302–328. https://doi.org/10.1080/15434303.2016.1246552

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515–535. https://doi.org/10.1177/0265532210368717

Birkel, P., & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss [How Concordant are Teachers' Essay Scorings? A Replication of Rudolf Weiss' Sudies]. *Psychologie in Erziehung Und Unterricht*, *49*(3), 219–224.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Chamberlain, S., & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology*, *42*(4), 665–675. https://doi.org/10.1111/j.1467-8535.2010.01062.x

Chernikova, O., Heitzmann, N., Fink, M.C. et al. (2019). Facilitating Diagnostic Competencies in Higher Education—a Meta-Analysis in Medical and Teacher Education. *Educ Psychol Rev, 32,* 157–196. https://doi.org/10.1007/s10648-019-09492-2

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing 1. *Educational Research and Evaluation*, *13*(5), 401–434. https://doi.org/10.1080/13803610701728311

Culham, R. (2003). *6+ 1 traits of writing: The complete guide*. New York: Scholastic Inc.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38–61. https://doi.org/10.1016/j.asw.2008.12.003

Driscoll, D. P., Avallone, A. P., Orr, C. S., & Crovo, M. (2010). *Writing framework for the 2011 National Assessment of Educational progress*. Washington, DC: National Assessment Governing Board, US Dept. of Education.

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills-Modified. *School Psychology Review*, *30*(1), 33–49.

European Commission (2008). Multilingualism - an asset for Europe and a shared commitment. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:ef0003

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: NY: Academic Press.

Flor, M., Futagi, Y., Lopez, M., & Mulholland, M. (2015). Patterns of misspellings in L2 and L1 English: A view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies*, *6*, 107–132. https://doi.org/10.15845/bells.v6i0.811

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, *71*(3), 328–338. https://doi.org/10.1037/0022-0663.71.3.328

Hamp-Lyons, L. (1991). *Assessing Second Language Writing in Academic Contexts*. Chestnut St., Norwood: Ablex Publishing Corporation.

Heitzmann, N., Fischer, F., & Fischer, M. R. (2018). Worked examples with errors: When self-explanation prompts hinder learning of teachers diagnostic competences on problem-based learning. *Instructional Science*, *46*(2), 245–271. https://doi.org/10.1007/s11251-017-9432-2.

Herppich, S., Praetorius, A.-K., Förster, N., Karst, K., Leutner, D., Behrmann, L., . . . Südkamp, A. (2017). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*. (76), 1–13. https://doi.org/10.1016/j.tate.2017.12.001

Huot, B. (1996). Toward a new theory of writing assessment. *College composition and communication*, *47*(4), 549-566. https://doi.org/10.2307/358601

Hyland, K. (2008). *Second language writing*. New York: Cambridge University Press. https://doi.org/10.1017/S0261444808005235

Jansen, T., Vögelin, C., Machts, N., Keller, S., & Möller, J. (2019). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen [The Student Inventory ASSET for judging students performances in the subject English: Three experimental studies on effect of text quality and student names]. *Psychologie in Erziehung Und Unterricht*, *66*(4), 303–315. https://doi.org/10.2378/peu2019.art21d

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education*, *97*, 103216. https://doi.org/10.1016/j.tate.2020.103216

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen [The student inventory: how student characteristics bias teacher judgments]. *Zeitschrift Für Erziehungswissenschaft*, *18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5

Kaufmann, E. (2020). How accurately do teachers judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe: Theorie, Prozessgestaltung, Empirie*. Tübingen: Gunter Narr Verlag.

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

Lovorn, M. G., & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teachers. *Practical Assessment, Research & Evaluation*, *16*(16), 1–18.

Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal*, *4*(4), 375–385.

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: AQA Centre for Education Research and Policy.

Murphy, K. R., & Reynolds, D. H. (1988). Does true halo affect observed halo? *Journal of Applied Psychology*, *73*(2), 235–238. https://doi.org/10.1037/0021-9010.73.2.235

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing*, *15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication*, *1*(4), 446–458. https://doi.org/10.1177/0741088384001004004

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

Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice*, *28*(2), 2–8. https://doi.org/10.1111/j.1745-3992.2009.00142.x

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, *44*(1), 57–84. https://doi.org/10.1002/tea.20163

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159-179. https://doi.org/10.1080/02602930801956059

Scannell, D. P., & Marshall, J. C. (1966). The effect of selected composition errors on grades assigned to essay examinations. *American Educational Research Journal*, *3*(2), 125–130.

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

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, *2*(1), 3–17. https://doi.org/10.1016/1060-3743(93)90003-L

Urhahne, D., & Wijnia, L. (2021). A Review on the Accuracy of Teacher Judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Vögelin, C., Jansen, T., Keller, S., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgments 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., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: an analysis of teacher comments. *The Language Learning Journal.* Advance online publication. https://doi.org/10.1080/09571736.2018.1522662

Weigle, S. C. (2002). *Assessing Writing.: Cambridge Language Assessment Series*. Cambridge: CUP.

Weir, C. (1988). The specification, realization and validation of an English language proficiency test. In Hughes A. (Ed.), *Testing English for university study. ELT documents 127* (pp. 45–110). London: Modern English Publications in association with The British Council.

Wind, S. A., Stager, C., & Patil, Y. J. (2017). Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments. *Assessing Writing*, *34*, 1–15. https://doi.org/10.1016/j.asw.2017.08.003

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

Zimmermann, F., Möller, J., & Köller, O. (2018). When students doubt their teachers' diagnostic competence: Moderation in the internal/external frame of reference model. *Journal of Educational Psychology*, *110*(1), 46–57. https://doi.org/10.1037/edu0000196.

6. Appendix: Analytic scales for argumentative Essays

Frame of essay: Introduction and conclusion

4 - Effective introduction with “hook” and “thesis statement”; effective conclusion summarizing main arguments

3 - Mostly effective introduction with either “hook” or “thesis statement”; mostly effective conclusion summarizing main arguments

2 - Introduction and/or conclusion identifiable but only partly effective

1 - Both introduction and conclusion not clearly identifiable or mostly ineffective

Body of essay: Internal organization of paragraphs

4 - Paragraphs are well-organized and coherent throughout

3 - Paragraphs are mostly well-organized and coherent

2 - Paragraphs are partly well-organized and coherent

1 - Paragraphs are not well-organized and incoherent

Support of arguments

4 - Author uses a variety of different examples to support her/his argument and fully explains their relevance to the topic

3 - Author uses different examples to support her/his argument and mostly explains their relevance to the topic

2 - Author uses a few examples to support her/his argument and partly explains their relevance to the topic

1 - Author uses repetitive examples to support her/his argument and their relevance to the topic is mostly unclear

Spelling and punctuation

4 - Author uses mostly correct spelling and punctuation

3 - Author uses mostly correct spelling and punctuation, with few distracting errors

2 - Author uses partly correct spelling and punctuation, with some distracting errors

1 - Author uses partly correct spelling and punctuation, with many distracting errors

Grammar

4 - Author uses a variety of complex grammatical structures, few grammar mistakes

3 - Author uses some complex grammatical structures, grammar mostly correct

2 - Author uses few complex grammatical structures, grammar partly correct

1 - Author uses few or no complex grammatical structures, grammar mostly incorrect

Vocabulary

4 - Author uses sophisticated, varied vocabulary throughout

3 - Author mostly uses sophisticated, varied vocabulary

2 - Author partly uses sophisticated, varied vocabulary, sometimes repetitive

1 - Author uses little sophisticated, varied vocabulary, often repetitive

Overall task completion

4 - Text fully conforms to the conventions of an argumentative essay, thus fully completing the task

3 - Text mostly conforms to the conventions of an argumentative essay, thus mostly completing the task

2 - Text partly conforms to the conventions of an argumentative essay, thus partly completing the task

1 - Text does not conform to the conventions of an argumentative essay, thus not completing the task