Linguistic Features of Formative Feedback on ESL Argumentative Writing: Comparing Pre-Service and Experienced Teachers

Cristina Vögelin
University of Basel

Stefan Daniel Keller
University of Basel

This experimental study investigated pre-service and experienced teachers' formative feedback responding to upper-secondary English as a Second Language (ESL) argumentative essays. It examined differences in feedback quality and linguistic features regarding teaching experience and text quality (high/low). We developed holistic criteria of effective formative feedback based on empirical findings in order to rate comments by 26 experienced and 41 pre-service teachers. Natural language processing tools were then applied to evaluate linguistic features of these comments. Results indicate that teachers provided more high-quality feedback to stronger essays than to weaker texts. No significant difference was found between pre-service and experienced teachers in terms of feedback quality. Further, comment length and absence of negative adjectives seem to predict feedback quality. Implications for research and practice are discussed.

Introduction

Providing feedback to students is an essential task of teaching English as a second or foreign language (ESL/EFL) writing (K. Hyland & Hyland, 2006a) and it can be a powerful influence on students' learning and motivation (Ferris, et al., 1997; Hattie & Timperley, 2007). Feedback is particularly crucial for acquiring and developing composition skills in process-based classrooms which focus on students' learning processes and areas to improve (K. Hyland & Hyland, 2006b; Shute, 2008). Effective feedback should thus identify where learners are in their learning process, where to proceed, and how to achieve their next goal (Hattie & Timperley, 2007; Parr & Timperley, 2010). Overall, the literature suggests that it is important for ESL teachers to acknowledge the range of needs, as well as strengths, in their students' writing and present them in a structured way (Ferris et al., 2011). Such high-quality feedback, however, is often unrealistic due to large class sizes, workloads, and formal requirements by institutions (Anson & Anson, 2017). Teachers report difficulties in finding a balance between positive and negative comments as well as a balance between content and form related aspects (Anson & Anson, 2017; Junqueira & Payant, 2015; Montgomery & Baker, 2007; Stern & Solomon, 2006). Many empirical studies have shown

Issues in Applied Linguistics ISSN 1050-4273

Vol. 22 No. 1, 1-22

that teachers' comments are predominantly negative (Connors & Lunsford, 1993; Read, Francis, & Robson, 2005; Stern & Solomon, 2006; Vögelin et al., 2018) and often focus on lower-order aspects, such as word choice, grammar or spelling, instead of higher-order concerns, such as content and discoursal issues (Anson & Anson, 2017; Lee et al., 2018; Montgomery & Baker, 2007). However, there is also evidence that teachers respond more to higher-order concerns than to formal criteria, such as spelling mistakes, when responding to student writing (Dixon & Moxley, 2013; Vögelin et al., 2018) and discrepancies have been reported when distinguishing weaker and stronger student writing (Cohen, 1987; Cumming, et al., 2002; Ferris et al., 1997). While feedback directed at weaker students predominantly focused on lower-order concerns, teachers attended to higher-order concerns in their feedback to stronger students. Negative here indicates that "the learner's utterance lacks veracity or is linguistically deviant" (Ellis, 2009, p. 3) and intends to promote a revision or correction. Thus, corrective feedback is one type of negative feedback.

Further, the effectiveness of teacher feedback on ESL writing for students' writing development has been extensively discussed in applied linguistics (Ene & Upton, 2018). Most studies report the effectiveness of teacher commentary (Biber et al., 2011; Ene & Upton, 2018; Ferris & Hedgcock, 2005; K. Hyland & Hyland, 2006b). At the same time, these studies also differ widely in how they analysed students' uptake from feedback (Goldstein, 2016; Guénette & Lyster, 2013; F. Hyland, Nicolás-Conesa, & Cerezo, 2016).

While a considerable body of literature has focused on feedback evaluating university students and their instructors (Cho et al., 2006; Dixon & Moxley, 2013; Ene & Upton, 2018; Ferris, 1994; F. Hyland & Hyland, 2001; Stern & Solomon, 2006), there are only a few studies examining (upper-) secondary ESL students and their teachers (Lee, 2007; Vögelin et al., 2018). Yet, as feedback can influence students' motivation and learning considerably (Hattie & Timperley, 2007), it is particularly important to investigate teachers' feedback practices also at secondary level. This study addresses this research gap by analyzing teachers' responses to ESL argumentative essays written by upper-secondary learners in Switzerland and Germany. It is relevant since argumentative essays are a crucial component of the upper-secondary ESL curriculum in both countries (Fleckenstein et al., 2020). Additionally, there is a need for further studies investigating the link between teacher factors, such as experience or subject knowledge, and the quality of formative feedback (Goldstein, 2004, 2016). This study presents a novel approach by developing holistic criteria for the quality of formative feedback based on the large body of previous empirical research findings. These criteria are then used to classify teachers' comments and to determine feedback quality. Based on these rated comments, we examine the influence of teaching experience (pre-service vs. experienced) on the quality of feedback. In addition, this study analyses the presence/absence of four selected linguistic features in teachers' comments with the help of natural language processing (NLP) tools and investigates whether selected linguistic features predict feedback quality.

Effective Formative Feedback

With the rise of the process approach in writing in the U.S., the importance of feedback was increasingly recognized by researchers and practitioners of EFL writing instruction from the 1970s onwards (Black & Wiliam, 2009; Ferris, 2003; F. Hyland et al., 2016). This approach – based on the cognitive theory of writing by Flower and Hayes (1981) – highlights the

importance of multiple drafts, feedback, and revisions during the process of writing. Perceiving writing as a process rather than a product gave further rise to formative feedback, which intends to support students in their writing process by identifying strengths as well as areas for improvement (Black & Wiliam, 2009; Ferris, 2003; Popham, 2009; Sadler, 1989). It is thus assessment *for* learning and stands in contrast to summative feedback – assessment *of* learning – which evaluates students' final writing product. Assessment for learning is aimed at promoting learning and engaging students in their learning process (Black & Wiliam, 2009).

It has been well demonstrated that effective feedback includes both encouraging and constructive criticism (Ferris, 2014). A balance between positive and negative comments implies that teachers do not simply list students' weaknesses, yet instead combine perceived areas for improvement with positive comments motivating students to revise their writing (Anson & Anson, 2017; Stern & Solomon, 2006). Effective positive feedback includes praise related to students' effort, performance, and engagement (Hattie & Timperley, 2007), in contrast to premature and gratuitous praise which can confuse and discourage students (K. Hyland & Hyland, 2006b). With the exception of K. Hyland and Hyland (2006b), who found more praise than criticism in teachers' written feedback, a range of studies suggested that teachers' comments consisted predominantly of negative aspects (Connors & Lunsford, 1993; Read et al., 2005; Stern & Solomon, 2006; Vögelin et al., 2018).

Studies further showed that effective feedback focuses on both content and form (Biber et al., 2011; Ene & Upton, 2018; Ferris, 2014), addressing a range of textual issues such as organization, content, and language (Ferris et al., 2011). Biber, Nekrasova, and Horn (2011) found that feedback providing a balanced account of aspects relating to both content and form led to greater learning gains of ESL students than feedback focusing only on form. Nevertheless, previous studies have reported that teachers focus more on local features, such as grammar and spelling (Connors & Lunsford, 1988; Ferris, 2003; Ferris et al., 2011; Goldstein, 2016; F. Hyland, 2003; Moxley, 1989, 1992; Parr & Timperley, 2010; Sommers, 1982; Stern & Solomon, 2006; Zamel, 1985), even though they described their approach to writing as comprehensive and global (F. Hyland, 2003). Additionally, previous research has shown that teachers' feedback differs depending on students' proficiency levels (Cohen, 1987; Ene & Upton, 2014; Ferris et al., 1997). For example, Ferris et al. (1997) reported that "teachers take a more collegial, less directive stance when responding to stronger students, while focusing more on surface-level problems with weaker students" (p. 175). Similarly, Cumming et al. (2002) found that ESL/EFL raters attended more extensively to rhetoric and ideas than language concerns in ESL compositions which they scored highly. Further, Ene and Upton (2014) showed teacher feedback decreased as student proficiency levels increased. This is problematic since feedback focusing on language accuracy has been proven to be insufficient for students' writing development and is unlikely to trigger their cognitive processing, in contrast to deeper-level feedback (Parr & Timperley, 2010). Dixon and Moxley (2013), however, found that instructors mainly focused on higher-order concerns, such as rhetoric, in contrast to lower-order concerns, such as grammar, regardless of the quality of the student's composition. Similarly, Vögelin et al. (2018) showed that teachers responded more to content-specific criteria than to formal criteria when responding to upper-secondary ESL essays.

Moreover, effective feedback includes text-specific comments (Ferris, 2014). Text-specific comments refer only to selected areas that are important for students' writing

performances concerning the particular task, instead of providing a comprehensive analysis of the students' weaknesses. This focused feedback appeared to be the most effective feedback regarding students' revisions (Ene & Upton, 2018; Ferris, 1997; Ferris & Hedgcock, 2005). For example, D. Ferris (1997) found that longer and text-specific comments led to more successful revisions by ESL tertiary students. Further, students appreciated comments that refer to specific problems and goals (Conrad & Goldstein, 1999; Goldstein, 2004; Hattie & Timperley, 2007), in contrast to generic comments without any reference to their individual strengths and weaknesses.

Effective feedback includes strategies for revision (Conrad & Goldstein, 1999; Goldstein, 2004; Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Stern & Solomon, 2006). For example, Hattie and Timperley (2007) reported the highest effect sizes for studies in which students received task-specific feedback and information about how to improve their writing. Moreover, Conrad and Goldstein (1999) showed that students struggled with comments that did not include an explicit strategy for revision and "either did not attempt revision or revised unsuccessfully in response to such comments" (p. 187). Concerning corrective feedback, Ene and Upton (2018) found that direct and explicit feedback was more effective than indirect feedback, especially for students at lower proficiency levels. Yet, indirect feedback was similarly effective as direct feedback when it was accompanied by meta-linguistic feedback explaining the next steps for revision.

Lastly, effective feedback is purposeful and meaningful (Hattie & Timperley, 2007). This principle highlights a core objective of effective feedback, namely identifying learners' current position in the learning process and supporting them to achieve their next goal. With a clear purpose in mind, teachers are able to create opportunities for students to close the gap in their learning development.

Possible reasons for ineffective feedback practices might be a shortcoming of institutional support or teacher training (Goldstein, 2016). Numerous teachers acquire their ability to give effective feedback only through practice (F. Hyland et al., 2016). Thus, it is central for teacher training programs to incorporate this essential aspect of diagnostic competence systematically in the curriculum. Further, L2 writing research has found that writing teachers often struggle with time constraints (Guénette & Lyster, 2013; Junqueira & Payant, 2015; Lee, 2009), and that there is a mismatch between teachers' beliefs and practices (e.g., Cumming, 1990; Ferris, 2014; Junqueira & Payant, 2015; Montgomery & Baker, 2007).

Teaching Experience and Providing Written Feedback

Differences in teaching experience and knowledge influence teachers' ability to provide quality feedback to students (Parr & Timperley, 2010). Teachers' knowledge encompasses an understanding of how writing works to achieve a communicative goal, support students in their learning process, and scaffold their learning (Jones & Moreland, 2005; Parr & Timperley, 2010). In the classic definition by Shulman (1986), professional knowledge is conceived as *content knowledge, pedagogical knowledge,* and *pedagogical content knowledge*. While content knowledge refers to teachers' subject matter knowledge per se, pedagogical content knowledge describes subject matter knowledge for teaching and includes, for example, the understanding to adapt particular topics according to learners' abilities and interests (Shulman, 1987). In addition to knowledge, experience is an essential component of teachers'

professionalism following the expert-novice paradigm (Gruber & Stöger, 2011). While experts possess extensive knowledge and experience regarding domain-specific tasks and great success in problem-solving and efficiency, novices need practice and experience to reach expert level (Gruber & Stöger, 2011). Furthermore, experts can draw on previous experiences to deal with new problems and typically focus on the deep structure of problems, in contrast to pre-service teachers who typically deal with the surface structure of problems. Among many competences, experts exhibit the ability to express themselves in a precise and sophisticated manner.

With regard to empirical studies, Rinnert and Kobayashi (2001) found that experienced EFL students and non-native English teachers were largely concerned with clarity, logical connections, and organization, while inexperienced EFL students concentrated mainly on content in written compositions. Cumming et al. (2002) showed that ESL/EFL raters predominantly focused on language-related aspects, while native English raters paid equal attention to language, rhetoric, and ideas overall. Thus, this study shows that ESL/EFL raters attended more to form-related aspects of student writing regardless of their experience. Cho, Schunn, and Charney (2006) investigated comments from undergraduate students, graduate students, and a subject matter expert and showed that comments decreased in length depending on experience. Further, studies from the field of writing assessment indicated that experienced teachers judged student texts stricter than novices (Barkaoui, 2010; Jansen, et al., 2021; Rinnert & Kobayashi, 2001). We conclude that teacher education has effects on teachers' feedback practices (Junqueira & Payant, 2015), yet training in responding to student writing needs to take contextual factors, such as institutional ideology or attitude, as well as local practices into consideration (Lee et al., 2018).

Automated Analysis of Linguistic Features in Teacher Feedback

An increasing amount of digital research tools allow extensive, fast, and efficient analyses of texts (Dixon & Moxley, 2013). These tools enable a systematic and efficient examination of phenomena occurring in natural language and thus provide the opportunity to observe patterns of language, e.g., in student writing, which otherwise would not be discernible by researchers (Hyatt, 2005). In contrast to student writing, written teacher feedback typically consists of short comments composed in full sentences or keywords. As these short texts mostly consist of opinion expressions and evaluative language, they resemble the genre of *product reviews*. In order to examine opinions expressed in reviews automatically, *sentiment analysis* has been conducted in numerous studies in the field of natural language processing and computer linguistics. Sentiment analysis describes the extraction of opinions, feelings, and emotions from natural language texts by employing text-mining techniques (Crossley, et al., 2017; Liu, 2015; Ren & Hong, 2017). It is able to evaluate the polarity of sentences, characteristics, or entire comments. While sentiment analysis has predominantly been applied to product reviews, educational discourse is an emerging field for the appliance of sentiment analysis (Crossley et al., 2017). Only recently, researchers have started to analyse sentiments in education using machine learning and natural language processing techniques (Rani & Kumar, 2017). For instance, Rajput et al. (2016) applied several text analytics methods to students' responses to open-ended questions in a course evaluation. A study by Crossley et al. (2017) verified that negative and positive reviews can be classified based on a number of lexical features related to sentiment. They found that adjectives were the most predictive feature of

positive as well as negative texts. These studies show the potential of digital tools to investigate textual features of teacher feedback.

Methodology

Research Questions

This paper aims to analyze pre-service and experienced secondary level teachers' written formative feedback on student writing with regard to its quality and four linguistic features. The following research questions guide this study:

1) Are there differences between pre-service and experienced teachers' formative feedback in terms of its quality?
2) Does student text quality influence the quality of teacher feedback?
3) Are there differences between pre-service and experienced teachers' formative feedback in terms of word count, lexical sophistication and the presence of negative adjectives as well as positive adjectives?
4) Do word count, lexical sophistication, negative adjectives, and positive adjectives predict high-quality feedback?

In line with previous research, we expect that participants provide more high-quality feedback responding to stronger student essays (Cohen, 1987; Cumming et al., 2002; Ferris et al., 1997). We further expect that experienced teachers write longer comments (Cho et al., 2006) using more sophisticated language (Gruber & Stöger, 2011). Last, we expect experienced teachers to phrase comments more negatively based on previous findings in the field of writing assessment (Barkaoui, 2010; Jansen et al., 2021; Rinnert & Kobayashi, 2001).

Student Texts

Structuring and presenting arguments coherently in a written text is an important learning objective in upper-secondary ESL classrooms in Switzerland and Germany (Brupbacher et al., 2008; KMK, 2012). Thus, the genre of argumentative essays was chosen for this study, and four authentic argumentative essays were selected from the Measuring English Writing at Secondary Level (MEWS) corpus of 906 ESL learner texts (Keller, 2016). The research project MEWS examined the writing competences of Swiss and German baccalaureate students at grade 11. Students were asked to write an independent essay answering the following TOEFL iBT® writing prompt: "Do you agree or disagree with the following statement? Television advertising directed toward young children (aged two to five) should not be allowed" (Rupp et al., 2019). Each text was scored by the e-Rater® and two human raters resulting in a human-human-machine (HHM) score ranging from 0 to 5 (Rupp et al., 2019). For this study, two texts with an HHM score of 3 and two texts with an HHM score of 4 were selected. The chosen texts were between 267 and 355 words long.

Rating Scale

In order to identify high-quality feedback in our sample, we developed a holistic rating scale based on previous findings on effective feedback (Anson & Anson, 2017; Biber et al., 2011; Conrad & Goldstein, 1999; Ene & Upton, 2018; Ferris, 1997, 2014; Goldstein, 2004,

2006; Hattie & Timperley, 2007; F. Hyland & Hyland, 2001; K. Hyland & Hyland, 2006a; Nicol & Macfarlane-Dick, 2006; Parr & Timperley, 2010; Stern & Solomon, 2006). The rating scale is given in Appendix A. Empirical findings on effective feedback were reviewed, structured, and summarized into five principles and were then included as descriptors for each level of the holistic rating scale. These descriptors were similarly worded to ensure consistent criteria ("is" – "is mostly" – "is partly" – "is not"). The rating scale had four levels of quality ranging from *effective feedback*, *mostly effective feedback, partly effective feedback,* to *mostly ineffective feedback.* After several rounds of adaptation and applying the criteria to comments from a pilot study, the authors decided on a final version of the rating scale.

Selected NLP Tools

The analysis of linguistic features in teachers' comments was conducted with the *Tool for Automatic Analysis of LExical Sophistication* (TAALES) and the *Sentiment Analysis and Social Cognition Engine* (SEANCE). TAALES measured lexical sophistication by calculating word range which refers to the number of texts in a corpus in which this word occurs (Crossley & Kyle, 2018). Words occurring in fewer contexts are generally more sophisticated (Kyle, Crossley, & Berger, 2018). The selected lexical sophistication indices are deduced from the Brown corpus, the British National Corpus, the Corpus of Contemporary American English, and the SUBTLEXus (Kyle et al., 2018). The four single lexical sophistication scores were averaged to one component score in the analysis. The freely available text analysis tool SEANCE was used to examine the polarity of comments (Crossley et al., 2017). SEANCE bases its analysis on several sentiment dictionaries, such as the General Inquirer (GI) lists (Stone, Dunphy, Smith, Ogilvie, & Associates, 1966), and polarity lists by Hu and Liu (2004). The tool analyses the polarity of a text by classifying text segments into positive or negative affect (Crossley et al., 2017). The tool also includes a negation feature, which ignores a target word in a particular category if it identifies a negation word in the three words preceding the target word. The reversion of polarity for negated sentences is an important component of accurate sentiment analysis. Besides a considerable number of other indices, SEANCE provides 20 component scores that combine similar micro features into larger macro features, and which offer a more manageable alternative for a simple exploration of sentiments (Crossley et al., 2017). We included the positive adjectives as well as the negative adjectives component scores since they proved to be the most predictive feature of the polarity of a review. *Beautiful, good,* and *amazing* are examples of positive adjectives, and *bad, poor,* and *terrible* are examples of negative adjectives. It is important to note that a negative score of negative adjectives indicates a positive comment overall.

Participants

A total of 83 pre-service teachers training for lower- and upper-secondary level and experienced secondary level teachers participated in the study. Both lower- and upper-secondary level teachers were included since their training is similar, consisting of a solid grounding in English linguistics and literature with added courses in teaching methodology and general education studies. Both types of teachers, furthermore, are expected to respond to argumentative student writing at some point in the relevant curricula. However, 15 participants were excluded since they did not compose written feedback to all four student texts. One

experienced teacher was further excluded since her teaching experience was limited to primary school level. While primary school teachers complete a general pedagogical education, secondary teachers receive an intensive education on the subject at universities and teacher education colleges. The remaining 67 participants consisted of $N = 26$ experienced teachers and $N = 41$ pre-service teachers. The age of participants ranged from 22 to 74 years, with a mean of 37.06 years ($SD = 7.10$). Participants were not equally divided by gender (71.6% female), which corresponds to the gender distribution at Schools of Education (Federal Office for Statistics, 2014). The majority of participants (74.6%) had (Swiss) German as their mother tongue, followed by English (9.0%) and various other languages (16.4%). Most participants reported that their English proficiency was equivalent to a C2 (46.3%) following the Common European Framework level (Council of Europe, 2001). The remaining participants described their English proficiency equivalent to a C1 (34.3%), B2 (6.0%), or had English as their L1 (13.4%). Pre-service teachers ($N = 41$) were students of higher education taking seminars at universities in Switzerland and Germany. Their age ranged from 22 to 36 years, with a mean of 25.27 years ($SD = 2.68$). On average, they had already completed 8.08 semesters ($SD = 2.52$) at university and reported little experience at teaching at upper-secondary level outside of their training ($M = 0.13$ years; $SD = 0.51$). Experienced teachers ($N = 26$) participated in the study on a voluntary basis and received a small remuneration. Their age ranged from 29 to 74 years with a mean of 48.85 years ($SD = 11.52$). The majority of teachers had a teaching degree for upper-secondary level (51.9%). Other teaching degrees included tertiary level (37.0%), lower-secondary level (11.1%), primary school (11.1%), and adult education (11.1%). Several teachers had more than one teaching degree. The average teaching experience at upper-secondary level was 11.63 years ($SD = 9.62$) and at other levels was 14.26 years ($SD = 10.28$). Thus, we can assume that the participants in this sample are able to provide level-specific feedback and are familiar with the writing requirements for upper-secondary level.

Procedure

In a computer-based assessment tool called Student Inventory ASSET (SIA), participants first received background information on the student texts: the assessment context, the writing task, learners' proficiency level, and age. Second, they were introduced to holistic and analytic rating scales and could read four student texts without assessing them to obtain a first impression. Third, participants assessed the four texts in randomized order using both scales. Each student essay was presented on the left-hand side in a split screen, while the rating scale appeared on the right-hand side. Fourth, participants were asked to fulfil the following task: "Assume that the student is going to revise and edit this text at least one more time before it is finalized. Please give the student some feedback for revision". Thus, participants wrote only one end-comment per text without providing marginal comments or in-text corrections. This procedure resulted in four written comments per participant. Last, participants were asked to answer background questions regarding their language proficiency, their teaching degree, and teaching experience.

Data Collection and Analysis

In total, 268 comments were analyzed. All comments were written in English, with the exception of one participant whose comments were translated from German to English. The

form of comments ranged from full sentences to keywords. In order to categorize the comments according to their quality, we developed a coding scheme based on the holistic rating scale for effective formative feedback, including anchor examples for the different levels. Table 1 displays two examples of effective feedback and mostly ineffective feedback.

Table 1
Examples of Feedback Quality

Feedback quality	Examples
Effective feedback	*"A very nice start to your essay! You've done an impressive job of finding facts and quotes to support your arguments. However, try to follow the structure we've discussed in class. You could also make a spelling checklist of words you often get wrong and use this before handing in your final."*
Mostly ineffective feedback	*"Good structure, partly good argumentation, unfortunately not so good, one could argue better, but all in all not bad! Please keep it up, you'll make it. The text would certainly be above average under a somewhat different task."*

Two trained raters used this rubric to analyze and score the quality of the teacher comments. The raters consisted of one author and a master's student with experience in coding language samples. The rater training encompassed jointly rating teacher comments from a pilot study in order to familiarize oneself with the coding scheme. Then, the two raters independently coded 40 training comments and, in case of disagreement, adjudicated upon a final score. After this training session, uncertainties, as well as the coding scheme, were discussed and specified. Finally, the two raters separately rated a representative 25% random selection of the data (68 teacher comments). The inter-rater reliability was $\kappa=0.78$, which is deemed substantial (Landis & Koch, 1977). The remaining data were thus equally divided and coded by each rater separately.

We conducted two-way repeated measures analysis of variance (ANOVA) with contrasts to investigate possible differences between pre-service and experienced teachers' comments in terms of feedback quality and student text quality. The dependent variable was feedback quality, the between-factor was teachers' experience (low/high), and the within-subjects factor was the text (four different student texts). The same analysis was chosen to examine whether pre-service and experienced teachers' formative feedback differs with regard to word count, lexical sophistication, negative adjectives, and positive adjectives. We refrained from conducting multivariate analyses due to the small sample size. A multiple linear regression was calculated to predict feedback quality based on word count, negative adjectives, positive adjectives, and lexical sophistication.

Results

Ranging between two to 304 words in length, the 268 comments had a mean length of M = 62.01 (SD = 43.98). The first research question asked whether there were differences between pre-service and experienced teachers' formative feedback in terms of its quality. The two-way repeated-measures ANOVA showed no significant effect for the variable experience ($F(1, 65) = 2.27$, $p = .14$). Therefore, pre-service and experienced teachers did not significantly differ in their ability to provide high-quality feedback. Table 2 displays means and standard deviations of feedback quality and linguistic features grouped by teaching experience.

Table 2

Means and Standard Deviations of Feedback Quality and Linguistic Features

	Experience	Low text quality		High text quality		Total	
		M	SD	M	SD	M	SD
Feedback quality	Pre-service	2.32	0.84	2.55	0.97	2.32	0.98
	Experienced	1.94	0.65	2.35	0.80		
Word count	Pre-service	65.63	34.13	58.21	31.75	62.01	43.98
	Experienced	60.92	49.97	63.37	56.41		
Lexical sophistication	Pre-service	1468.19	107.38	1474.65	142.11	1420.99	188.85
	Experienced	1319.10	205.30	1363.83	160.35		
Negative Adjectives	Pre-service	-0.29	0.57	-0.59	0.65	-0.40	0.88
	Experienced	-0.05	0.77	-0.62	0.56		
Positive Adjectives	Pre-service	0.31	0.32	0.42	0.40	0.31	0.48
	Experienced	0.08	0.31	0.37	0.34		

Feedback quality: 1 = mostly ineffective; 4 = effective

There were, however, significant effects for the four different student texts ($F(3, 195) = 6.06$, $p = .001$, $\eta^2 = .09$). No significant interaction effect between the different texts and experience was found ($F(3, 195) = .65$, $p = .59$).

The second research question investigated whether student text quality influenced the quality of teacher feedback. Results of the within-subjects contrasts revealed significant differences between low and high text quality ($F(1, 65) = 12.27$, $p = .001$, $\eta^2 = .159$). Thus, the quality of formative feedback directed towards stronger student essays was higher than feedback directed towards weaker student essays. No significant differences between the two stronger texts ($F(1, 65) = 2.47$, $p = .12$) and the two weaker texts ($F(1, 65) = 2.15$, $p = .15$) were found.

The third research question examined whether there were differences between pre-service and experienced teachers' formative feedback in terms of word count, lexical sophistication and the presence of negative adjectives and positive adjectives. Concerning

comment length, results of the two-way repeated-measures ANOVA showed significant effects for different texts ($F(3, 195) = 5.03$, $p < .01$, $\eta^2 = .072$) and a significant interaction effect between different texts and experience ($F(3, 195) = 4.08$, $p < .01$, $\eta^2 = .059$). Results of the within-subjects contrasts revealed significant differences between the two stronger texts ($F(1, 65) = 17.10$, $p < .001$, $\eta^2 = .208$). Thus, comments responding to the two stronger texts differed significantly in word length. In contrast, no significant differences were found between low and high text quality ($F(1, 65) = 0.96$, $p = .33$) and between the two weaker texts ($F(1, 65) = 0.03$, $p = .86$).

Concerning the interaction between texts and teaching experience, results showed a significant effect between the two weaker texts and experience ($F(1, 65) = 6.88$, $p = .01$, $\eta^2 = .096$). While pre-service teachers provided longer comments, experienced teachers wrote shorter comments responding to the second weaker text, in comparison to the first weaker text. Overall, no significant effect for experience with regard to word count was found ($F(1, 65) = 0.00$, $p = .98$). Results of the two-way repeated-measures ANOVA with the independent variable lexical sophistication did not yield significant effects for different texts ($F(3, 195) = 2.44$, $p = .07$), nor an interaction effect between texts and experience ($F(3, 195) = 1.25$, $p = .29$). Tests of between-subjects effects, however, yielded significant effects for experience ($F(1, 65) = 15.57$, $p = .001$, $\eta^2 = .193$). It must be noted that lexical sophistication was measured by word range and lower values indicate writing that is more sophisticated. Therefore, pre-service teachers employed less sophisticated language in their comments than experienced teachers.

Regarding the negative adjectives component score, results indicated significant effects for different texts ($F(3, 195) = 6.15$, $p = .001$, $\eta^2 = .086$), but no significant interaction effect for texts and experience ($F(3, 195) = .59$, $p = .62$). Contrasts showed significant effects for low and high text quality ($F(1, 65) = 16.97$, $p < .001$, $\eta^2 = .207$). Hence, comments addressed to weaker student texts were more negative than comments addressed to stronger texts. No significant differences between the two weaker texts ($F(1, 65) = .73$, $p = .40$), between the two stronger texts ($F(1, 65) = .48$, $p = .49$) and no significant effect for the variable experience was found ($F(1, 65) = .82$, $p = .37$). Lastly, results of the two-way repeated-measures ANOVA with Greenhouse-Geisser correction for the positive adjective component score indicated significant effects for different texts ($F(2.42, 157.46) = 4.44$, $p < .01$, $\eta^2 = .064$). No interaction effect between texts and teaching experience was found ($F(2.42, 157.46) = 1.14$, $p = .33$). Contrasts showed significant effects between low and high text quality ($F(1, 65) = 13.64$, $p < .001$, $\eta^2 = .173$). Comments addressed to stronger texts were more positive. No significant effects were found for the two stronger texts ($F(1, 65) = 0.06$, $p = .81$) and for the two weaker texts ($F(1, 65) = 0.28$, $p = .60$). Further, results showed a significant effect for experience ($F(1, 65) = 4.00$, $p = .05$, $\eta^2 = .058$), indicating that pre-service teachers provided more positive comments.

The fourth research question asked whether word count, lexical sophistication, negative adjectives, and positive adjectives predicted high-quality feedback. Table 3 displays the correlation matrix of quality of teachers' feedback and four linguistic features.

Table 3
Correlations between Feedback Quality and Linguistic Features

Measure	1	2	3	4	5
1) Feedback quality	-				
2) Word count	.66 ***	-			
3) Lexical sophistication	.26 ***	.24 ***	-		
4) Negative Adjectives	-.16 **	-.03	-.16 **	-	
5) Positive Adjectives	-.01	-.12 *	.18 **	-.44 ***	-

Note: * $p < .05$ ** $p < .01$ *** $p < .001$.

It is worth noting that the length of comments correlated highly with the quality of formative feedback. Results of the multiple linear regression analysis showed that word count, negative adjectives, positive adjectives, and lexical sophistication explained 45.2% of variance ($F(4, 267) = 56.11$, $p < .001$). As Table 4 displays, word count significantly predicted the quality of formative feedback, indicating that longer comments were rated to be more effective.

Table 4
Multiple Linear Regression Predicting Quality of Feedback

Model	*β (standardized)*	*SE*	*p*-value
Word Count	**.631**	**.001**	**.000**
Negative Adjectives	**-.133**	**.056**	**.009**
Positive Adjectives	-.004	.106	.944
Lexical Sophistication	.090	.000	.061
Adjusted $R^2 = .452$			

Further, results showed that the presence of negative adjectives significantly predicted the quality of written comments, indicating that the quality of comments with a negative polarity was lower. No significant results were found for positive adjectives and lexical sophistication.

Discussion

The purpose of this study was to examine pre-service and experienced teachers' written formative feedback with regard to its quality, linguistic features and student text quality. First, the analysis of pre-service and experienced teachers' formative feedback showed no significant difference in terms of its quality. This finding is contrary to the belief that differences in experience influence teachers' ability to provide quality feedback to students (e.g., Gruber & Stöger, 2011; Parr & Timperley, 2010; Shulman, 1986). The study further investigated whether

there are differences between pre-service and experienced teachers' formative feedback regarding student text quality. Results displayed significant differences between stronger and weaker student texts – thus, the quality of feedback responding to stronger student essays was higher than feedback addressed to weaker texts. This is in line with previous research reporting that teachers primarily focus on surface-level problems and thus provide insufficient feedback when addressing weaker students (Cohen, 1987; Cumming et al., 2002; Ferris et al., 1997).

The third research question focused on linguistic features in formative feedback and possible differences of these features in terms of teaching experience and student text quality. No significant difference in comment length concerning teaching experience was found. This result stands in contrast to previous findings stating that comment length increases with teaching experience (Cho et al., 2006). Yet, in this study, teachers were free to choose their writing format as they were only instructed to provide formative feedback to the student. Hence, the comments included both full sentences as well as keywords, which might have skewed the results of comment length. Concerning lexical sophistication, there is evidence suggesting that pre-service teachers employ less sophisticated language in their comments, which corresponds to the expert-novice paradigm (Gruber & Stöger, 2011). No significant difference in lexical sophistication was found for text quality, indicating that teachers do not differ their language responding to weaker or stronger student essays. Results further showed that comments responding to weaker texts were more negative than comments addressed to stronger texts. In return, comments responding to stronger texts exhibited a more positive polarity. While this finding might not be surprising, it indicates that the SEANCE component scores were able to measure relevant differences in teachers' comments. Further, another important finding was that pre-service teachers' comments were more positive than experienced teachers' ones. This result corresponds with previous findings in quantitative writing assessment which showed that experienced teachers judge student texts more strictly (Barkaoui, 2010; Jansen et al., 2021; Rinnert & Kobayashi, 2001). It further conforms to the descriptive analysis of the most frequent bigrams in this study that disclosed more positive bigrams in pre-service teachers' comments.

Last, this study examined how well word count, lexical sophistication, negative adjectives, and positive adjectives predicted feedback quality. The data showed that comment length significantly predicted the quality of formative feedback. Thus, longer comments are rated as more effective than shorter ones. A possible explanation for this finding might be the length-sensitive holistic criteria for formative feedback, which were employed in this study based on empirical findings. The criteria state that effective feedback includes both encouraging comments and constructive criticism, is purposeful and meaningful, focuses on both content and form, includes a variety of text-specific comments, and includes explicit strategy for revision (e.g., Anson & Anson, 2017; Ene & Upton, 2018; Ferris, 1997; Hattie & Timperley, 2007; K. Hyland & Hyland, 2006b; Parr & Timperley, 2010; Stern & Solomon, 2006). To fulfil the majority of these criteria, comments needed to reach a certain length. Further, the study has found that the absence of negative adjectives predicted feedback quality. Thus, the quality of comments with a negative polarity was lower. A possible explanation is that comments with fewer negative adjectives were rated higher in terms of quality because they were perceived as more constructive and encouraging for students.

Limitations and Implications for Practice

There are several limitations to this study that should be discussed. First, this study investigated an experimental assessment situation that differs markedly from an authentic classroom situation. Writing in a classroom context implies a unique combination and interplay of factors with regard to the institution, teachers, and students. These include the writing course, the teacher-student relationship, or teachers' expectations about their students (Goldstein, 2004; Lee et al., 2018). The context in which teachers are embedded can influence their feedback practices considerably (Lee et al., 2018). In an authentic classroom, teachers know their students – their strengths, weaknesses as well as their prior performances – and this personal relationship is reflected in their feedback. Further, teachers evaluate and comment on students' written responses with a distinct goal of the writing task in their minds. This mental representation of the desired output, at least in theory, helps teachers to write comments that best move students toward improvement. In this experimental study, participants were isolated from all interpersonal dimensions of feedback and did not possess a deep understanding of the writing task in the larger trajectory of their students' academic progress. Our results should therefore be interpreted with caution and cannot be generalized directly to a real classroom context. In contrast to the multi-faceted classroom context, however, our experimental research design allows the analysis of single determinants of teachers' written responses with an attempt to minimize interference of other confounding variables.

Second, this study focused on teachers' comments without measuring students' improvement, which does not necessarily follow even after effective feedback (Hattie & Timperley, 2007; Sadler, 1989). Yet, we believe that our study marks a beginning of studies investigating formative feedback on ESL argumentative essays in a Swiss and German context since it identifies concrete linguistic aspects of high-quality feedback. The study also included the background variable teaching experience. Such studies are relevant since they shed light on current feedback practices in schools, (pre-service) teachers' ability to provide feedback, and how feedback should be formulated in order to be effective. Future studies could further examine formative feedback in the context of authentic writing classrooms. Further, the small sample size of $N = 67$ participants is rather limiting within the complex research design of an experimental study with four texts of differing text qualities. Thus, we included only four linguistic features in our analysis. It would be interesting to evaluate more linguistic features in a larger data set or even compile a context-specific sentiment dictionary for teacher comments on ESL learner essays.

Although the sample in this study was relatively small, and it is difficult to generalize the results, we can draw several implications which are discussed in relation to practice and research. Overall, the findings suggest that – following our holistic criteria for formative feedback, which is based on numerous empirical studies – both pre-service and experienced teachers on average only provided partly high-quality feedback. This might be due to a discrepancy between empirical findings of effective feedback and teachers' beliefs as well as practices. In this context, the holistic scale of high-quality feedback developed in this study could be beneficial as it could be implemented and discussed in teacher education in the future, addressing teachers' difficulties to give well-balanced feedback (Anson & Anson, 2017; Junqueira & Payant, 2015; Montgomery & Baker, 2007; Stern & Solomon, 2006). Together

with introducing such a holistic scale in teacher education, specific workshops should focus on aspects such as the importance of feedback strategies to provide well-balanced feedback, or ways of providing feedback to large classes and numerous courses. Insecurities as well as constraints from institutional circumstances and requirements, which might conflict with teachers' expectations and practices, should be addressed explicitly – and with a suitable set of tools – in teacher education.

Lastly, this study found that pre-service secondary level teachers provided more positive comments than experienced teachers. While the importance of encouraging comments for students' development has been well demonstrated in research, experienced teachers might feel less inclined to accompany their corrective suggestions with positive comments due to time pressure, overwhelming workload, and their students' expectations. Many ESL students believe that comments concerning errors and form-related aspects of their writing are a critical part of writing instruction (Lee et al., 2018; Montgomery & Baker, 2007). The results of this study thus hold important clues for the development of training and workshops aimed specifically at secondary teachers. One goal in such a workshop could be for teachers to discover whether their own feedback is mostly positive or negative, and whether they tend to focus more on higher or lower-order concerns. They could learn about the effects of both types of feedback on students' development based on recent findings from empirical literature while considering the implications of this research on their own feedback practices. Further, trainings and workshops could exemplify effective comments and therefore raising teachers' awareness of linguistic features in high-quality comments. By showing them prototypes of effective comments, teachers could further practice producing such comments both in spoken and written form. To create authentic learning opportunities, they could first engage in such an exercise 'at leisure' and later practice it under time pressure.

While this study's research lays the groundwork for these practical implications, it needs to be expanded. More research investigating ESL pre-service and experienced teachers' perceptions and beliefs on formative feedback is needed to address the issue of how to improve teachers' knowledge of – and motivation for – formative feedback (Ferris, 2014; Junqueira & Payant, 2015; Montgomery & Baker, 2007). Further studies should evaluate the effectiveness of such feedback circles in actual classrooms at different levels in order to examine their usefulness in contextualized environments.

References

Anson, I. G., & Anson, C. M. (2017). Assessing peer and instructor response to writing: A corpus analysis from an expert survey. *Assessing Writing*, *33*, 12–24.

Barkaoui, K. (2010). Do ESL Essay Raters' Evaluation Criteria Change with Experience? A Mixed-Methods, Cross-Sectional Study. *TESOL Quarterly*, *44*(1), 31–57.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis*. Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31.

Brupbacher, B., Jucker, A. H., König, E., Roth, M., & Straumann, B. (2008). Englisch. In A. Hsgym (Ed.), *Hochschulreife und Studierfähigkeit—Zürcher Analysen und Empfehlungen zur Schnittstelle* (pp. 88–96).

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on Writing—Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication*, *23*(3), 260–294.

Cohen, A. D. (1987). Studying Learner Strategies: Feedback on Compositions. *PASAA*, *17*(2), 29–38.

Connors, R. J., & Lunsford, A. A. (1988). Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research. *College Composition and Communication*, *39*(4), 395–409.

Connors, R. J., & Lunsford, A. A. (1993). Teachers' Rhetorical Comments on Student Papers. *College Composition and Communication*, *44*(2), 200–223.

Conrad, S. M., & Goldstein, L. M. (1999). ESL Student Revision after Teacher-Written Comments: Text, Contexts, and Individuals. *Journal of Second Language Writing*, *8*(2), 147–179.

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

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, *38*, 46–50.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*, 803–821.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal*, *86*(i), 67–96.

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*, 241–256.

Ellis, R. (2009). Corrective Feedback and Teacher Development. *L2 Journal*, *1*(1), 3–18. https://doi.org/10.5070/l2.v1i1.9054

Ene, E., & Upton, T. A. (2014). Learner uptake of teacher electronic feedback in ESL composition. *System*, *46*, 80–95. https://doi.org/10.1016/j.system.2014.07.011

Ene, E., & Upton, T. A. (2018). Synchronous and asynchronous teacher electronic feedback and learner uptake in ESL composition. *Journal of Second Language Writing*, *41*, 1–13.

Federal Office for Statistics, B. F. S. (2014). *Educational achievement (Bildungsabschlüsse)*. Neuchâtel.

Ferris, D. (1994). Lexical and Syntactic Features of ESL Writing by Students at Different Levels of L2 Proficiency. *TESOL Quarterly*, *28*(2), 414–420.

Ferris, D. (1997). The Influence of Teacher Commentary on Student Revision. *TESOL Quarterly*, *31*(2), 315–337.

Ferris, D. (2003). *Response to Student Writing—Implications for Second Language Students*. New York: Routledge.

Ferris, D. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing*, *19*, 6–23.

Ferris, D., Brown, J., Liu, H., Eugenia, M., & Stine, A. (2011). Responding to L2 Students in College Writing Classes: Teacher Perspectives. *TESOL Quarterly*, *45*(2), 207–234.

Ferris, D., & Hedgcock, J. S. (2005). *Teaching ESL composition: Purpose, process, and practice*. Mahwah, NJ: Lawrence Erlbaum.

Ferris, D., Pezone, S., Tade, C., & Tinti, S. (1997). Teacher Commentary on Student Writing: Descriptions & Implications. *Journal of Second Language Writing*, *6*(2), 155–182.

Fleckenstein, J., Keller, S. D., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing, 43*(100420). https://doi.org/10.1016/j.asw.2019.100420

Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *National Council of Teachers of English*, *32*(4), 365–387.

Goldstein, L. M. (2004). Questions and answers about teacher written commentary and student revision: Teachers and students working together. *Journal of Second Language Writing*, *13*, 63–80.

Goldstein, L. M. (2006). Feedback and revision in second language writing: Contextual, teacher, and student variables. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing—Contexts and Issues* (pp. 185–205). Cambridge University Press.

Goldstein, L. M. (2016). Making use of teacher written feedback. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of Second and Foreign Language Writing* (pp. 407–430). De Gruyter.

Gruber, H., & Stöger, H. (2011). Experten-Novizen-Paradigma. In E. Kiel & K. Zierer (Eds.), *Unterrichtsgestaltung als Gegenstand der Wissenschaft*. Baltmannsweiler: Schneider Verlag Hohengehren.

Guénette, D., & Lyster, R. (2013). Written Corrective Feedback and Its Challenges for Pre-Service ESL Teachers. *The Canadian Modern Language Review*, *69*(2), 129–153. https://doi.org/10.3138/cmlr.1346

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112.

Hu, M., & Liu, B. (2004). *Mining and Summarizing Customer Reviews* (W. Kim & R. Kohavi, Eds.). ACM Press.

Hyatt, D. F. (2005). 'Yes, a very good point!': A critical genre analysis of a corpus of feedback commentaries on Master of Education assignments. *Teaching in Higher Education*, *10*(3), 339–353.

Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System*, *31*, 217–230.

Hyland, F., & Hyland, K. (2001). Sugaring the pill—Praise and criticism in written feedback. *Journal of Second Language Writing*, *10*, 185–212.

Hyland, F., Nicolás-Conesa, F., & Cerezo, L. (2016). Key issues of debate about feedback on writing. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of Second and Foreign Language Writing* (pp. 433–452). De Gruyter.

Hyland, K., & Hyland, F. (2006a). *Feedback in Second Language Writing—Contexts and Issues*. Cambridge University Press.

Hyland, K., & Hyland, F. (2006b). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101.

Jansen, T., Vögelin, C., Machts, N., Keller, S. D., & Köller, O. (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

Jones, A., & Moreland, J. (2005). The importance of pedagogical content knowledge in assessment for learning practices: A case-study of a whole-school approach. *The Curriculum Journal*, *16*(2), 193–206.

Junqueira, L., & Payant, C. (2015). 'I just want to do it right, but it's so hard': A novice teacher's written feedback beliefs and practices. *Journal of Second Language Writing*, *27*, 19–36.

KMK. (2012). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch / Französisch) für die Allgemeine Hochschulreife*. Berlin.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.

Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning? *Assessing Writing*, *12*, 180–198.

Lee, I. (2009). Ten mismatches between teachers' beliefs and written feedback practice. *ELT Journal*, *63*(1), 13–22. https://doi.org/10.1093/elt/ccn010

Lee, J. J., Vahabi, F., & Bikowski, D. (2018). Second Language Teachers' Written Response Practices: An In-House Inquiry and Response. *Journal of Response to Writing*, *4*(1), 34–69.

Liu, B. (2015). *Sentiment Analysis—Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Montgomery, J. L., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing*, *16*, 82–99. https://doi.org/10.1016/j.jslw.2007.04.002

Moxley, J. M. (1989). Responding to Student Writing: Goals, Methods, Alternatives. *Freshman English News*, *17*, 3–11.

Moxley, J. M. (1992). Teachers' Goals and Methods of Responding to Student Writing. *English Faculty Publications*, *20*(1), 17–33.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.

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

Popham, W. J. (2009). Assessment Literacy for Teachers: Faddish or Fundamental? *Theory Into Practice*, *48*, 4–11.

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-Based Sentiment Analysis of Teachers' Evaluation. *Applied Computational Intelligence and Soft Computing*, 1–12.

Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. *Computer*, 36–43.

Read, B., Francis, B., & Robson, J. (2005). Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, *30*(3), 243–262.

Ren, G., & Hong, T. (2017). Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach. *Sustainability*, *9*. https://doi.org/10.3390/su9101765

Rinnert, C., & Kobayashi, H. (2001). Differing Perceptions of EFL Writing among Readers in Japan. *The Modern Language Journal, 85*(ii), 189–209.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(4), 4–14.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–21.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189.

Sommers, N. (1982). Responding to Student Writing. *College Composition and Communication*, *33*(2), 148–156.

Stern, L. A., & Solomon, A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing*, *11*, 22–41.

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., & Associates. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

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.* https://doi.org/10.1080/09571736.2018.1522662

Zamel, V. (1985). Responding to Student Writing. *TESOL Quarterly*, *19*(1), 79–101.

Appendix A

Holistic Criteria for Formative Feedback

4 – Effective feedback
• Includes both encouraging comments and constructive criticism • Is purposeful and meaningful • Focuses on both content and form • Includes a variety of text-specific comments • Includes explicit strategy for revision
3 – Mostly effective feedback
• Includes some encouraging comments, some constructive criticism • Is mostly purposeful and meaningful • Focuses mostly on both content and form • Includes some text-specific comments, few generic comments • Includes mostly explicit strategy for revision
2 – Partly effective feedback
• Includes few encouraging comments, little constructive criticism • Is partly purposeful and meaningful • Focuses partly on content, mostly on form • Includes few text-specific comments, mostly generic comments • Includes partly explicit strategy for revision
1 – Mostly ineffective feedback
• Includes no encouraging comments and constructive criticism. Demotivating criticism or premature and gratuitous praise • Is not purposeful and meaningful • Focuses on form and errors alone • Includes few or no text-specific comments, entirely generic comments • Includes no explicit strategy for revision

Cristina Vögelin is a Ph.D. candidate in Educational Sciences at University of Basel. Her research interests include second language acquisition, language assessment and corpus linguistics.

Stefan D. Keller is a professor of Teaching and Learning of English Language and its disciplines at the School of Education, University of Applied Sciences and Arts Northwestern Switzerland. He is deputy director of the Institute for Educational Sciences, University of Basel.