Assessing Writing 39 (2019) 50–63

Contents lists available at ScienceDirect

Assessing Writing

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

The influence of lexical features on teacher judgements of ESL argumentative essays

Cristina Vögelin[a,*], Thorben Jansen[b], Stefan D. Keller[a], Nils Machts[b], Jens Möller[b]

[a] Institute for Educational Sciences, University of Basel, Hofackerstrasse 30, 4132 Muttenz, Switzerland
[b] Institute for Psychology of Learning and Instruction, Kiel University, Olshausenstrasse 75, 24118 Kiel, Germany

ARTICLE INFO

Keywords:
Lexical diversity
Lexical sophistication
Teacher education
Teacher judgements
Second language writing
Writing assessment

ABSTRACT

Numerous studies have examined the relationship between lexical features of students' compositions and judgements of text quality. However, the degree to which teachers' judgements are influenced by the quality of vocabulary in students' essays with regard to their assessment of other textual characteristics is relatively unexplored. This experimental study investigates the influence of lexical features on teachers' judgements of English as a second language (ESL) argumentative essays. Using analytic and holistic rating scales, English pre-service teachers (N = 37) in Switzerland assessed four essays of different proficiency levels in which the levels of lexical diversity and sophistication had been experimentally varied. Coh-Metrix software was used to manipulate the level of lexical diversity, as measured by MTLD and D, and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) software was used to obtain differing levels of lexical sophistication, as measured by word range. The results suggested that texts with greater lexical diversity and sophistication were assessed more positively concerning their overall quality as well as the analytic criteria 'grammar' and 'frame of essay'. The implications of this study for classroom practice and teacher education are discussed.

1. Introduction

The genre of argumentative essays is a key component of learning and teaching English as a second language (ESL) in many European countries, particularly at upper-secondary level (Keller, 2013; Zemach & Stafford-Yilmaz, 2008; Zhu, 2001). This competence is particularly relevant in Switzerland and Germany, as English is the lingua franca of science, business and higher education in these countries (Keller, 2013; Porsch & Köller, 2010). National guidelines for Swiss and German baccalaureate schools state that learners should be able to "gather and structure information and present it coherently in written form" (1995, EDK, 1994). This involves learners becoming familiar with different ways of structuring arguments and supporting them with effective rhetorical devices (Brupbacher, Jucker, König, Roth, & Straumann, 2008; KMK, 2012).

From the importance of the genre itself, it follows that assessing this type of writing is a central dimension of teachers' diagnostic competences, as their assessment directly influences student learning and course achievement (2007, Hamp-Lyons, 2016; Weigle, 2002; White, 2009). It plays an important role in giving information on students' educational progress (Baumert & Kunter, 2006; Schrader, 2013), and it forms the basis for decisions on grades, transfers to the next level and diplomas (Kronig, 2007). However,

* Corresponding author at: Cristina Vögelin, Institute for Educational Sciences, University of Basel, Hofackerstrasse 30, 4132 Muttenz, Switzerland.
E-mail addresses: cristina.voegelin@unibas.ch (C. Vögelin), tjansen@ipl.uni-kiel.de (T. Jansen), ste.keller@unibas.ch (S.D. Keller), nmachts@ipl.uni-kiel.de (N. Machts), jmoeller@ipl.uni-kiel.de (J. Möller).

https://doi.org/10.1016/j.asw.2018.12.003
Received 2 November 2017; Received in revised form 3 December 2018; Accepted 4 December 2018
Available online 13 December 2018

studies of teacher training in Switzerland and Germany suggest that teacher trainees are not sufficiently trained in assessing complex and multidimensional student compositions (Porsch, 2010) and do not feel sufficiently prepared to assess students' compositions when they enter the profession (Rauin & Meier, 2007). More research is needed to understand the subject-specific components of teacher judgements and to develop methods to prepare teachers for the assessment of writing (Alderson, Brunfaut, & Harding, 2014).

Following Sadler (1989), the term *assessment* here is used as "any appraisal (or judgement, or evaluation) of a student's work or performance" (p. 120), and it can be formative or summative in its function (Black & Wiliam, 2003; Bloom, Hastings, & Madaus, 1971; Scriven, 1967). Summative assessment summarizes the learning achievement and aims to grade and certify students. Its primary purpose is to generate grades, which will be the basis for subsequent certification (Bearman et al., 2016). Formative assessment provides information about the quality of students' responses to help them improve their performance and is aimed at supporting student learning. According to Sadler (1989), both summative and formative assessments of student work involve qualitative judgements made directly by a person. Hermeneutic in nature and common in a wide variety of subjects, such judgements are particularly salient in assessing writing, where student development is multidimensional, and learning cannot be conceptualized as a sequence of separate units of skills or knowledge (Sadler, 1989). Argumentative essays are the outcome of complex thinking, planning, and organizing processes. ESL students need to apply their linguistic skills as well as their rhetorical and strategic competences in writing (Bereiter & Scardamalia, 1987; Eckes, 2008; Flower & Hayes, 1981; Hyland, 2008). From the perspective of assessment, these texts are multileveled in that the characteristics salient for determining their quality relate to different levels: word level (e.g., spelling, lexis), sentence and paragraph level (e.g., syntax, cohesion) and text level (e.g., coherence, structure, and argumentation). To assess the quality of students' written compositions, teachers must either configure these levels in a holistic assessment or hold them apart in an analytic assessment. Therefore, they must possess a concept of writing quality appropriate for the task and substantiate their judgement by referring to relevant criteria. Additionally, teachers need to be able to recognize and describe fine performances, to indicate areas for improvement in poor performances and to apply judgement criteria according to the standards of the profession and relevant to a certain genre (Crusan, Plakans, & Gebril, 2016; Sadler, 1989).

Uninformed or biased judgement can lead to a student losing time, motivation, and confidence, or, in the worst case, it can lead to a student being assigned to an unsuitable educational track. In this study, we use the lens model (Brunswik, 1956) to illustrate the process of forming judgements as well as underlying influences on raters' judgements. Through a lens, which is shaped by various textual or environmental cues, teachers' judged achievement estimate of a student text deviates from the true achievement estimate, which cannot be observed. With regard to factors influencing teacher assessment, we can distinguish between construct-relevant and construct-irrelevant influences (Messick, 1994). In our study, all aspects of assessment directly connected to text characteristics, such as spelling, organization, and vocabulary, are considered construct-relevant factors. In contrast, background factors, such as students' gender, social status and ethnicity (not included in this study), are construct-irrelevant factors. The teacher's task in such an assessment is to judge the quality of the compositions and thus the underlying writing competences of the students, irrespective of personal background factors or social stereotypes (Kaiser, Südkamp, & Möller, 2016; Südkamp, Kaiser, & Möller, 2012). By extension, it would be inappropriate to take one aspect of an essay, such as the quality of vocabulary, as a basis for judging a different, unrelated aspect of the same essay, such as organization or argumentation.

The aim of this study is to explore the influence of lexical features on teachers' holistic and analytic assessment. We manipulated the vocabulary in ESL argumentative essays by varying the values of lexical diversity and sophistication, as measured by the programs Coh-Metrix and Tool for the Automatic Analysis of Lexical Sophistication (TAALES), and studied in detail how teachers' judgements were affected by differing levels of lexical features. Whereas there is an interplay of different factors in a real assessment situation, this study presents an experimental research design with the aim of investigating single determinants of teacher judgements without the interference of confounding variables. We believe that it is important that key factors influencing assessment processes are identified to improve classroom assessment and teacher education.

2. Theoretical background

2.1. The relation between text characteristics and writing quality

Numerous studies have explored the relationship between linguistic characteristics of student essays and writing quality (Crossley & McNamara, 2011, 2012; Crossley, Salsbury, & McNamara, 2011; Grant & Ginther, 2000; Guo, Crossley, & McNamara, 2013; Yu, 2010), showing that lexical features predict L2 writing quality assessed by expert raters and automated scoring tools in high-stakes assessment contexts. For example, Grant and Ginther (2000) analysed – among others – lexical features in 90 ESL learner essays written at three proficiency levels. The results revealed differences among the three proficiency levels: As proficiency increases, there is a steady increase in lexical specificity (type-token ratio, average word length), conjuncts, emphatics and amplifiers (Grant & Ginther, 2000). In another study, Guo et al. (2013) examined whether linguistic features (lexical sophistication, syntactic complexity, cohesion and basic text information) predict L2 writing proficiency in the integrated and independent writing tasks of the Test of English as a Foreign Language (TOEFL iBT). They found that lexical sophistication is a significant predictor for both tasks. The findings of these studies demonstrate that linguistic characteristics predict judged writing quality and thus are the premise for the manipulation of lexical features in the current study.

2.2. The relation between text characteristics and teacher judgements

In addition to the body of empirical work showing that lexical features predict writing quality assessed by expert raters and

automatic scoring tools, no studies – to our knowledge – investigate how the assessment of professional raters differs from that of prospective and experienced teachers. In work comparing levels of expertise in teachers, various studies found little or no difference between pre-service and experienced teachers' assessment of students' written compositions (Meadows & Billington, 2010; Royal-Dawson & Baird, 2009; Shohamy, Gordon, & Kraemer, 1992). A considerable body of research, however, investigates how text characteristics influence teachers' assessment of student texts. It is well demonstrated that pre-service and experienced teachers assign lower grades to essays containing mechanical errors (Cumming, Kantor, & Powers, 2002; Marshall, 1967; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966). For example, Scannell and Marshall (1966) showed that prospective teachers are affected by the quality of composition, which was varied by inserting punctuation, grammatical and spelling errors, even when they were explicitly instructed to grade on content alone. In a more recent study, Rezaei and Lovorn (2010) showed that well-written essays containing 20 structural, mechanical, spelling and grammar errors were assigned lower scores than texts without errors even in criteria relating solely to content, such as "understanding and synthesis of argument" (p. 27). Teachers failed to distinguish between formal errors and the independent quality of content in a student essay. This indicates a halo effect (Thorndike, 1920), which appears "when judgments of one rated characteristic influence judgments of other characteristics in a positive or negative direction" (Bechger, Maris, & Hsiao, 2010).

In German contexts, researchers have examined and identified substantial difficulties for teachers to judge students' L1 and L2 texts objectively (Birkel & Birkel, 2002; Ingenkamp & Lissmann, 2008). For example, a negative effect was reported when German teachers assessed their students' English texts with a high number of spelling mistakes (Birkel & Birkel, 2002). Learners with a deficit in formal language skills were in danger of being underestimated in other areas, such as the organization, argumentation or structure of an essay. Furthermore, previous research has shown that teachers attend more extensively to formal language issues in their assessment of student essays in comparison to other aspects, such as organization or structure, especially if they are non-native English speakers (Cumming et al., 2002; Porsch, 2010; Rezaei & Lovorn, 2010).

In ESL writing, Cumming et al. (2002) found that ESL raters tend to focus more on students' use of language than on ideas or argumentation in comparison to native-language raters. In a factor analysis, Porsch (2010) showed that the latent factor *learners' writing performance* explained more variance in lexis ($\lambda = .80$) and grammar ($\lambda = .83$) than in content ($\lambda = .51$) and organization ($\lambda = .64$). In addition to a considerable number of studies investigating the relationship between text characteristics and writing proficiency, there is a lack of systematic research in the particular field of English essays and vocabulary (Ferris, Pezone, Tade, & Tinti, 1997; Olinghouse & Wilson, 2013). No study to our knowledge has investigated experimentally the influence of vocabulary on other dimensions of teacher judgements. Olinghouse and Wilson (2013) examined the role of vocabulary across story, informative and persuasive text and noted a lack of studies exploring the relationship between persuasive writing quality and vocabulary. Two key features of vocabulary in persuasive writing are lexical diversity (Jarvis, 2013) and lexical sophistication (Nation & Webb, 2011), which are also important indicators of L2 academic achievement (Daller, Milton, & Treffers-Daller, 2007). The following section provides more information about the theoretical background of these two components of lexical richness and their relationship to rated writing proficiency.

2.3. Lexical diversity

Lexical diversity, or lexical variation, is typically defined as the range and variety of vocabulary in a learner's language use (Malvern, Richards, Chipere, & Durán, 2004; McCarthy & Jarvis, 2007; Yu, 2010). A number of studies have found links between language proficiency and lexical diversity, demonstrating that higher-rated essays are associated with greater lexical diversity (Crossley, Kyle, Allen, Guo, & McNamara, 2014; Crossley, Salsbury, & McNamara, 2014; Engber, 1995; Jarvis, 2002; Linnarud, 1986; Yu, 2010). Engber (1995) reported significant correlations between lexical variation including lexical errors ($r = .45$) and excluding errors ($r = .57$) and a holistic score of writing quality. Essays with a more varied vocabulary were evaluated more positively. Additionally, Crossley, Salsbury et al. (2014) reported that lexical diversity ($r = .70$), collocation accuracy ($r = .91$) and word frequency ($r = -.61$) are highly correlated with holistic scores of lexical proficiency. Taken together, these studies suggest that lexical diversity is an important indicator of writing proficiency (Crossley & McNamara, 2012), although the reported effect sizes are usually low to moderate (Crossley & McNamara, 2012; Engber, 1995; Grant & Ginther, 2000; Jarvis, Grant, Bikowski, & Ferris, 2003).

2.4. Lexical sophistication

Lexical sophistication, which is typically defined as the percentage of 'advanced' or less frequent words in a text sample, is another important factor in learners' writing proficiency (Laufer & Nation, 1995; Lu, 2012; Read, 2000; Yu, 2010). Using the *Lexical Frequency Profile,* Laufer and Nation (1995) showed that more proficient L2 learners used significantly fewer high-frequency words in their texts. Other studies showed similar findings with regard to word frequency counts based on different corpora (Crossley & McNamara, 2012; Crossley, Salsbury et al., 2014; Crossley, Cobb, & McNamara, 2013; Cumming et al., 2005; Engber, 1995; Guo et al., 2013; McNamara, Crossley, & McCarthy, 2010; Read, 2000). Thus, as the level of language proficiency increases, the degree of lexical sophistication increases as well. A recent study by Kyle and Crossley (2016) measured lexical sophistication in argumentative L2 writing and showed a significant relationship between word range and essay quality (*BNC Written Range All Words* explained 16.7% of variance in essay scores). Hence, higher-rated essays tend to include words with a more restricted word range containing more specific and domain appropriate words. Nevertheless, further research is needed to systematically test whether word range is a robust predictor of essay quality across different domains and language levels (Kyle & Crossley, 2016).

3. Purpose

The purpose of this experimental study is to investigate the influence of the text quality and lexical features of argumentative essays on teacher judgements. Accordingly, we examine pre-service teachers' analytic and holistic assessment of writing quality. Furthermore, we explore whether the manipulation of lexical diversity and sophistication in ESL written compositions leads to halo effects concerning other analytic criteria, such as the frame of the essay, grammar, or the support of arguments. Consequently, the following research questions guide this study:

- To what degree are pre-service teachers able to differentiate between texts with different levels of overall text quality?
- To what degree does the level of lexical diversity and sophistication in ESL argumentative essays affect pre-service teachers' holistic and analytical judgement of these texts?

This experimental study contributes to research and practice for writing assessment in different ways. First, it extends research on the relationship between lexical features and teachers' judgements of argumentative essays. Second, it can be considered a high-stakes writing assessment in the context of Switzerland and Germany, since the English end-of-year grade at this level consists primarily of in-class argumentative essays. Third, it shows how the text analysis tools Coh-Metrix and TAALES provide suitable measures to manipulate the level of lexical features in students' written compositions. Fourth, this study adds to a more complete understanding of teacher judgements and the influencing factors in the assessment process. This information can be used in teacher training to prepare pre-service teachers for real assessment situations in their ESL classrooms.

4. Method

4.1. Context

This study is part of the larger research project ASSET, "Assessing Students' English Texts" (co-founded by the national bodies of scientific research in Switzerland and Germany). Based on the heuristic model of teacher judgement accuracy by Südkamp et al. (2012), the project investigates the influence of teacher, student, test and judgement characteristics on teachers' judgements of ESL argumentative essays in a series of studies.

The experimental design of this study is based upon the *lens model* of Brunswik (1956), which describes the quality of rater judgements and the underlying processes (Wind, Stager, & Patil, 2017). According to this model, the quality of a rater judgement is constituted by the difference between the true achievement estimate, which cannot be observed, and the observed judged achievement estimate. The judged estimate is dependent on a variety of cues, which represent the lens through which raters infer the true achievement level. Cues can encompass domains, rubrics, essay features and others (Wind et al., 2017). In this study, cues are limited to essay features, such as the vocabulary, grammar, and frame of the essay and other analytic dimensions that influence a teacher's assessment of student compositions. By manipulating one characteristic of an ESL argumentative essay and thus experimentally varying the true achievement estimate of the student text, we explore the influence of vocabulary on the judged achievement estimates of other textual features. Whereas differing qualities of lexical diversity and sophistication are relevant for the assessment of *vocabulary*, the manipulation of vocabulary is a construct-irrelevant factor for the assessment of *frame of essay*, which is concerned with the introduction and conclusion of an argumentative essay.

4.2. Instruments

4.2.1. Measuring lexical diversity

Measuring lexical diversity is a rather controversial linguistic issue, and a number of different measures have been introduced to date (Durán, Malvern, Richards, & Chipere, 2004; Jarvis & Daller, 2013; Yu, 2010). One of the best-known measures is the type-token ratio (TTR), which is calculated by dividing the number of types (total number of different words) by the number of tokens (total number of words) in a language sample (Yu, 2010). TTR has been criticized, however, for its sensitivity to text length, as its values decrease with increasing sample size due to a higher degree of word repetition (Lu, 2012; Malvern et al., 2004). A number of studies have attempted to create an improved measure, yet there is currently little consensus among researchers on the strongest measure of lexical diversity, as each measure captures only a limited aspect of the concept (McCarthy & Jarvis, 2013).

In this study, we used the two tools D and Measure of Textual Lexical Diversity (MLTD) in combination, currently seen as the best available measurement of lexical diversity (McCarthy & Jarvis, 2010). The parameter D, calculated by *vocd*, models a family of ideal curves of TTR against tokens, and its value specifies the level of lexical diversity of the sample, representing the best fit curve for a language sample (Lu, 2012; McCarthy & Jarvis, 2013). Thus, a higher value of D indicates a lexically more diverse text. A more mathematical discussion of D is beyond the scope of this study (cf. Malvern et al., 2004; McCarthy & Jarvis, 2007 in details). Malvern et al. (2004) provided evidence that D is a reliable and valid index to measure lexical diversity. Furthermore, Yu (2010) demonstrated a significant correlation of $r = .33$ between D and a holistic assessment of students' writing performance as well as a significant correlation of $r = .48$ between D and their oral performance. Crossley, Salsbury et al. (2014) reported correlations ranging from $r = .8$ to $r = .9$ between holistic ratings of essay and D. As even D exhibits text length dependency, we chose only texts longer than 300 words and kept the text length constant at approximately 460 words in all essays used for the study. This is reported as the safest option for analysing lexical diversity (Treffers-Daller, 2013) and reduced the possibility of measurement errors due to text length.

The second measure for lexical diversity employed in this study was MTLD (McCarthy, 2005). It works as an index for lexical diversity that is not a function of text length (Jarvis, 2013). Briefly, MTLD calculates the mean length of word sequences used in a text to remain above a TTR value of .72 (McCarthy & Jarvis, 2010). Thus, a text with a high level of lexical diversity requires more words to reach this point of stabilization than a less diverse text. Studies have demonstrated that MTLD is an effective measure of lexical diversity and is even stronger than the measure D (Crossley, Salsbury et al., 2014; McCarthy & Jarvis, 2010, 2013; McNamara et al., 2010).

In this study, we used four learner texts, each with “high” or “low” lexical richness, yielding eight different text variations (c.f. Section 3.3.2.). To measure D and MTLD in these eight text variations, the computational tool Coh-Metrix 3.0 was used (Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix is a freely available online text analysis software primarily designed to assess text cohesion and includes 106 measures of text information, such as lexical diversity (McNamara, Graesser, McCarthy, & Cai, 2014). A number of studies have shown that the automated indices significantly predict holistic human ratings of lexical proficiency (Crossley & McNamara, 2011; Crossley, Cobb et al., 2013; Crossley, Salsbury, & McNamara, 2013; Crossley, Salsbury et al., 2014).

4.2.2. Measuring lexical sophistication

Lexical sophistication – the ratio between the number of sophisticated or advanced words and the total number of words – strongly relies on the researcher's definition of "sophisticated words" (Engber, 1995; Laufer & Nation, 1995; Linnarud, 1986). The concept is often based on corpus frequency counts, which disclose the relative difficulty of a lexical item (Kyle & Crossley, 2016; Laufer & Nation, 1995). However, Kyle and Crossley (2016) results indicate that word range and bigram frequency are better predictors of independent essay quality than word frequency, which is a common measure for lexical sophistication in several previous studies (Crossley, Cobb et al., 2013; Laufer & Nation, 1995; Linnarud, 1986). Therefore, this study measures lexical sophistication by word range values, which represent "the number of texts in a reference corpus in which a word occurs" (Kyle & Crossley, 2016, p. 14). Words with a high range value occur in numerous texts and various contexts throughout the corpus, whereas words with a low range value occur in a limited number of texts and contexts (Kyle & Crossley, 2016). Kyle and Crossley (2015) showed that range values are negatively correlated with analytic scores of lexical proficiency. With regard to holistic scores of writing proficiency and word range, Kyle and Crossley (2016) found a significant correlation between the index of word range and holistic writing proficiency scores in independent TOEFL writing samples. To calculate the word range of our four student texts, the freely available online tool TAALES was employed (Kyle & Crossley, 2015). The range indices by the same authors (2016) are deduced from the Brown corpus, British National Corpus (BNC), SUBTLEXus, and the Corpus of Contemporary American English (COCA). The selection of the four corpora ensured a sufficient range of contemporary technological words, which was needed since the learner compositions dealt with the topic of modern technology.

4.2.3. Analytic and holistic rating scales

To obtain objective and fair assessments, judgements of written compositions should be made with reference to reliable criteria such as assessment standards or rating scales (Sadler, 1989; Weigle, 2007). In this study, we included both a holistic rating scale and an analytic rating scale, which are the two most common and widely used types of rubrics (Crusan, 2010; Knoch, 2009; Rakedzon & Baram-Tsabari, 2017; Weigle, 2002). The reason for including both rating scales is two-fold. First, teachers in practice typically employ both holistic and analytic rating scales, referring to types of rubrics widely used in the field. On the one hand, the holistic approach focuses on the work as an entity and integrates the essential qualities of writing (Hamp-Lyons, 2003; Sadler, 1989). By evaluating the work as a whole, the assessor makes an entire assessment based on separate criteria in which “imperfectly differentiated criteria are compounded as a kind of gestalt and projected onto a single scale of quality” (Sadler, 1989, p. 132). On the other hand, the analytic approach enables teachers to assess the quality of student writing in separate criteria in a more systematic, differentiated way and provides students with more detailed information in their own achievement (Sadler, 1989). Thus, including both types of rubrics enhances the authenticity and practicality of the experiment as the intended application of this study is classroom assessment. The second reason relates to our research design: this experimental study investigates whether the influence of a linguistic feature is discernible on both types of scales to gain greater insight into assessment processes relevant to a school context.

Regarding teachers’ judgements of overall text quality, the well-validated NAEP rating scale with six levels using age- and grade-appropriate writing criteria was employed (Board, 2010). In a German-speaking context, the scale was previously used in a large-scale empirical assessment of writing in Germany (Knopp, Jost, Nachtwei, Becker-Mrotzek, & Grabowski, 2012). Previous research has shown that holistic ratings as single indices are insufficient since formal aspects, such as vocabulary and grammar, appear to have a stronger influence on the overall assessment compared to content and organization (Porsch, 2010). Moreover, holistic rubrics are not always easy to interpret since one predominant text dimension can distort the assigned score and outweigh other dimensions (Weigle, 2002).

In contrast, analytic rubrics facilitate an assessment including multiple scales, such as vocabulary and organization, thereby allowing more detailed and multi-faceted assessments of text (Weigle, 2002). They are particularly helpful for the assessment of texts written by second language learners, who tend to exhibit an uneven profile across different dimensions of writing, especially for untrained raters (Eckes, Müller-Karabil, & Zimmermann, 2016; Hamp-Lyons, 2003; Weigle, 2002). To address the genre-specific characteristics of argumentative essays at an upper-intermediate level (Hyland, 1990; Zemach & Stafford-Yilmaz, 2008), an analytic rubric was designed by adapting the Test in English for Educational Purposes (TEEP) by Weir (1988) and the 6 + 1 trait model by Culham (2003). The resulting rating scales contains seven dimensions: Frame of essay (introduction and conclusion), body of essay (organization of paragraphs), support of arguments, spelling and punctuation, grammar, vocabulary, and overall task completion. These seven dimensions are held to be particularly important in the context of teacher assessment (Coe, Hanita, Nishioka, & Smiley,

2011; Culham, 2003) and cover the salient characteristics of multileveled ESL student texts. Each aspect was divided into four levels with similarly worded descriptors ranging from “fully” – “mostly” – “partly” – “no(t)” (see Appendix).

The boundaries between characteristics in the analytic rubric are sometimes fuzzy, as argumentative texts are complex and multi-layered artefacts (Sadler, 1989). In the case of idioms, fixed expressions and similar lexical constructs, for example, it can be difficult to distinguish between grammar and vocabulary, since they carry both lexical as well as grammatical components (Lewis, 1993). Similarly, usage-based approaches to language learning sometimes accentuate the interconnection between vocabulary and grammar from an acquisition perspective (Staples & Reppen, 2016). From an assessment perspective and for purposes of diagnostic assessment, however, aspects such as grammar or lexical richness tend to be seen as separate categories (Council, 2001), and the distinction is often made by teachers in the classroom (Parr & Timperley, 2010). Furthermore, the distinction between vocabulary and grammar is appropriate for a study focussing on the assessment of specific L2 language subskills (Weigle, 2002). In particular, ESL learners tend to exhibit uneven learning profiles across different aspects of writing (Eckes et al., 2016; Weigle, 2002).

The rubric employed in this study was based on existing models yet was equally geared towards the relevant syllabus (Hawkey & Barker, 2004; Knoch, 2009; Rakedzon & Baram-Tsabari, 2017). To eliminate imprecise wording and ambiguities, the *a priori* created rubric underwent two rounds of piloting, revision and adaptation. Hence, the employed assessment rubrics represent both the underlying theory of the genre-specific characteristics for the chosen writing prompt and the developer’s conception of what skills and subskills are measured by the essay examination (Eckes et al., 2016; Meier, Trace, & Janssen, 2016; Rakedzon & Baram-Tsabari, 2017; Weigle, 2002).

4.3. Student essays

4.3.1. Selection for current study

In a first step, a teaching unit on argumentative writing was developed and implemented at an upper-secondary grammar school ("Gymnasium") in Basel, Switzerland. The students were in the 11th grade, had learnt English for 5 years and were upper-intermediate learners. Over four weeks, 15 ESL learners were taught techniques of argumentative writing with materials from the course book *Writers at Work - The Essay* (Zemach & Stafford-Yilmaz, 2008). The instruction was aligned with the dimensions of the analytic rubric used in this study. Therefore, students were taught not only how to structure and organize different elements of an argumentative essay but also how to sustain their arguments by providing different types of support and by using the appropriate type of language required by the genre. Moreover, the learners studied model texts of argumentative essays concerned with social and technological issues to be prepared for the final essay examination. In the final double lesson, all learners were asked to write a timed argumentative essay in 90 min to the following prompt: “Do you agree or disagree with the following statement? As humans are becoming more dependent on technology, they are gradually losing their independence”.

Experts from the School of Teacher Education (Basel) evaluated a total of 15 student texts together with the ASSET research team using the NAEP rating scale (Board, 2010). As a result, two 'stronger' and two 'weaker' texts, each of roughly equivalent overall quality, were chosen from the sample and adjusted to the same text length. The weaker texts exhibited levels of work that received a failing grade, i.e., considered insufficient by all experts with regard to the learning goals of the unit. 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. In a next step, the level of lexical diversity and sophistication was measured by Coh-Metrix and TAALES. Within the range of the data set, the MTLD, D and word range values were systematically varied, resulting in eight text variations, i.e., each of the four argumentative essays with a low level and a high level of lexical sophistication and diversity.

4.3.2. Manipulation of lexical features

Prior to analysis, spelling errors in all texts were corrected since Coh-Metrix and TAALES require exact spelling to generate accurate wordlists without incorrectly spelt words misrepresenting the vocabulary of a student (Hawkey & Barker, 2004). This study acknowledges the relationship between spelling and vocabulary as well as between lexical errors and writing quality (Read, 2000); however, this experimental design treated them as different characteristics of word knowledge and considered them as separate categories for text assessment. Any other errors in the four learner texts were retained.

From the dataset of 15 texts, four texts were chosen for variation, and their length was controlled for by the procedure of truncation since essay length relates to perceived writing proficiency (Wolfe, Song, & Jiao, 2016). The mean length for the four ESL argumentative essays was $M = 459.75$ ($SD = 3.96$). The level of lexical sophistication and diversity was then systematically lowered and increased in each of these four texts. Thus, the same text existed with high and low levels of lexical sophistication and diversity. This resulted in eight text variations: two texts with low overall text quality in two variations each (high/low quality of vocabulary) and two texts with high overall text quality in two variations each (high/low quality of vocabulary). To lower the level of lexical sophistication in the four texts, less sophisticated words were substituted for more sophisticated words following frequency word lists (The British National Corpus, 2007). For example, the verb *go away* replaced the verb *disappear* in the original sentence: “What would happen if all these devices that make our daily lifes [*sic*] so easy and comfortable just disappeared?”. To lower the level of lexical diversity, word repetitions were integrated and used to replace a diverse use of vocabulary, as this example illustrates:

- “As long as you see technology through this aspect, you will be happy to have **new technologies**.” (Original, own emphasis)
- “As long as you see technology through this aspect, you will be happy to have **innovations**.” (Manipulation, own emphasis)

The manipulated lexical items were partly taken from lexical items occurring in the four original texts. For example, one text with

Table 1
Lexical diversity – Data set.

	Lowest value	Highest value	*M (SD)*
MTLD	69.94	115.15	89.86 (13.64)
D	58.43	120.38	80.67 (19.02)

a low overall text quality exhibited many repetitions of the word *technology*, which was then incorporated in the text variations with a low level of lexical richness. Furthermore, explicit linking words, such as *additionally, furthermore* and similar expressions, were excluded from replacement to manipulate only lexical richness and not the categories *frame of essay* and *body of essay*.

Table 1 displays the values of lexical diversity of the data set, with its lowest and highest values, and Table 2 the level of lexical diversity of the four manipulated text variations. The degree of manipulation was limited to the existing range of lexical diversity and sophistication within the data set, since, for example, lexical diversity results in a poorer quality of language beyond a certain threshold of D (Jarvis, 2002). Whereas a higher score indicates greater lexical diversity, a lower score signifies a greater word range and thus greater lexical sophistication (see Table 3 and 4).

The resulting eight text variations included in the final experiment had a mean text length of 463.9 words (SD = 4.83). to account for the slightly varying D values due to the method of random sampling, the programme *vocd* was run 15 times for each text to obtain a final D (Yu, 2010). To present authentic learner texts to our participants, original spelling errors were re-integrated into the eight manipulated text variations. If words with spelling errors were adapted, the new words included similar spelling errors. Limitations associated with this procedure are discussed at the end of this article.

4.4. Participants

Participants were students of higher education taking a seminar on language skills for English teachers at a Swiss university (N = 37). The age of the participants ranged from 23 to 42 years with a mean of 29.5 years (SD = 5.1). Students were not equally divided by gender (81.1% female). This gender distribution corresponds to the high percentage of female students (85.3%) at Schools of Education in Switzerland (Federal Office for Statistics, 2014). With the exception of one student, all participants took the course as a requirement to obtain a teaching diploma. The majority of participants (75.7%) had English as their first subject. Second subjects included History, French, German, Politics, Art, Mathematics, Geography, Scandinavian studies, and others. With regard to their English proficiency, 5 participants had English as their L1, whereas the other students reported that their English proficiency was equivalent to a C2 (78.4%) or a C1 (8.1%) following the Common European Framework level (Council, 2001). A minority of students (35.1%) had already started their teaching training and had a mean of 3.8 weeks (SD = 2.5) of training in actual classrooms. Concerning their teaching experiences at secondary level outside of their training, 24.3% of participants had already taught for one whole semester or longer, and 21.6% reported teaching experience for single lessons within the context of a substitution. The majority, 51.4%, had no teaching experience at secondary level outside of their training. The overall sample thus consisted of pre-service teachers with high English language proficiency and low to middle teaching experience.

4.5. Procedure

This study employed the *Student Inventory ASSET* (SIA) adapted from the "student inventory" (SI) (Kaiser, Möller, Helm, & Kunter, 2015). In this computer-based assessment tool, participants first received background information on the school context and teaching situation from which the texts had originated: learners' proficiency level, age, and previous knowledge; the essay prompt; and the time limit. The introduction further included a description of the teaching unit that was implemented at the school prior to the writing prompt. Second, participants were introduced to both the holistic and analytic rating scales to familiarize themselves with the different descriptors. Whereas the scales themselves were in English, the task instructions were in German to guarantee maximum clarity. Third, they read the four essays without assessing them to obtain a first impression. Fourth, the participants assessed the four texts in a randomized order using the analytic and holistic rating scales. Each participant assessed four texts – two texts with low overall text quality as well as different levels of lexical richness (high/low) and two texts with high overall text quality as well as different levels of lexical richness (high/low). In a split screen, each student essay was presented on the left-hand side, and the rating scales were presented on the right-hand side (Fig. 1).

Participants were asked to select the competence level they thought was appropriate for the text following the presented descriptors. Lastly, participants filled in a background questionnaire regarding their level of education and language proficiency.

Table 2
Lexical diversity – Variation.

	Low lexical diversity				High lexical diversity			
	Text 1	Text 2	Text 3	Text 4	Text 1	Text 2	Text 3	Text 4
MTLD	72.04	72.58	74.20	79.53	100.97	101.78	105.07	107.56
D	83.55	83.77	86.08	79.75	110.61	105.67	105.64	108.27

Table 3
Lexical sophistication – Data set.

	Lowest value	Highest value	*M (SD)*
SUBTLEXus	6584.47	5484.02	6054.65 (336.23)
BNC	81.50	68.11	76.73 (3.63)
COCA	.68	.52	.60 (.04)
Brown Corpus	321.86	292.12	308.18 (9.74)

Table 4
Lexical sophistication – Variation.

	Low lexical sophistication Texts 1-4 *M (SD)*	High lexical sophistication Texts 1-4 *M (SD)*
SUBTLEXus	6378.16 (197.80)	5771.84 (356.60)
BNC	77.55 (5.17)	75.02 (3.02)
COCA	.62 (.03)	.57 (.03)
Brown Corpus	317.78 (6.39)	299.12 (10.70)

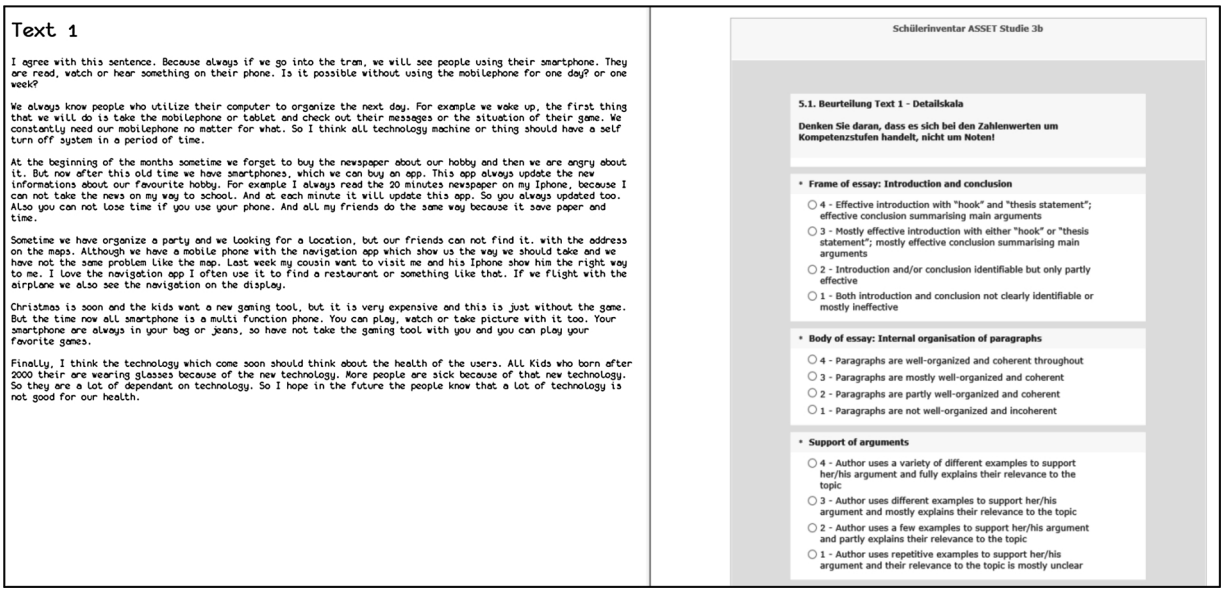

Fig. 1. Screenshot of the Student Inventory ASSET (SIA).

4.6. Analysis

To answer the research question, we used an experimental 2 × 2 design with two factors: overall text quality and vocabulary (lexical diversity and sophistication). The design was analysed with repeated-measures ANOVA (holistic scores) and a multivariate analyses of variance (analytic scores). This analysis assumes independence, sphericity and normal distribution of the data. While the first two assumptions were given by design, the assumption of a normal distribution could be violated due to the ordinal rating scales. However, the analysis of variance is robust to that violation since our study was a balanced design with more than 25 participants per condition (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). Furthermore, the skewness and kurtosis were in acceptable ranges for all dependent variables (see Table 5) and the raters used all scale levels. Our approach differs from recent developments in the estimation of multiple text parameters and rater effects through item response analyses (Eckes, 2015; Engelhard & Wind, 2018). We utilized a strictly experimental approach on the text level. As such, we varied the same texts through specific changes in the vocabulary. Consequently, it was our intention not to derive unknown text parameters from judgements but to explicate specific effects of the text variation on the judgements. Due to data collection on the computer, there were no missing values as participants could proceed only when they completed their assessment of four text variations. For each scale, inter-rater agreement was calculated using *ICC* (2,1) (Shrout & Fleiss, 1979). The relatively low ICC coefficients indicate substantial variation in the ratings of each of the eight texts in comparison to the variation in the ratings across the texts. However, the eight texts include two each of four different texts, with a slight variation in the vocabulary and only two differences in the overall text quality across the four texts. Thus, we can safely

Table 5
Descriptive statistics for rating scales.

	Min	Max	Mean (SD)	Skewness (SE)	Kurtosis (SE)	ICC
Holistic scale	1	6	3.67 (1.25)	.013 (.199)	−.530 (.396)	.199
Frame of essay	1	4	2.58 (.86)	.034 (.199)	−.673 (.396)	.252
Body of essay	1	4	2.51 (.82)	.102 (.199)	−.529 (.396)	.109
Support of arguments	1	4	2.64 (.84)	−.123 (.199)	−.549 (.396)	.178
Spelling	1	4	2.70 (.69)	−.241 (.199)	.004 (.396)	.189
Grammar	1	4	2.49 (.80)	.221 (.199)	−.438 (.396)	.367
Vocabulary	1	4	2.39 (.89)	.128 (.199)	−.705 (.396)	.322
Overall task completion	1	4	2.51 (.86)	−.042 (.199)	−.621 (.396)	.135

assume that the low ICCs are a result of the low overall variance of text quality, and we expect to see high ICCs only on the ratings of vocabulary.

5. Results

5.1. Teacher judgements of overall text quality

The result for the holistic score displayed the effects of text quality ($F(1,36) = 23.8$, $p < .001$, $\eta^2 = .398$). No significant interaction effect between quality and vocabulary with regard to the holistic scores was found ($F(1,36) = 3.3$, $p \geq .05$). Multivariate tests indicated significant effects for text quality and analytic scores ($F(7,30) = 11.7$, $p < .001$, $\eta^2 = .732$). No effects were detected for first-level interactions between quality and vocabulary with regard to the analytic scores ($F(7,30) = 1.7$, $p \geq .05$). For the significant multivariate effects, we conducted univariate post hoc tests. The first set of univariate tests indicated that texts with a high text quality were judged more positively with regard to all criteria than texts with a low text quality (see Table 6). The inter-rater agreement of the judgements on each scale varied between *ICC* (2,1) = .109 and *ICC* (2,1) = .367.

5.2. Teacher judgements of manipulated texts

Concerning the variable *vocabulary*, the result for the holistic score indicated a significant effect ($F(1,36) = 4.5$, $p < .05$, $\eta^2 = .110$). Multivariate tests indicated significant effects for vocabulary ($F(7,30) = 5.6$, $p < .001$, $\eta^2 = .556$). The second set of univariate tests disclosed that vocabulary had effects on the judgements of the criteria *frame of essay* ($F(1,36) = 2.7$, $p < .05$, $\eta^2 = .142$) and *grammar* ($F(1,36) = 3.6$, $p < .01$, $\eta^2 = .260$) (see Table 7). The results for the manipulation check showed an effect of vocabulary with regard to the criterion *vocabulary* ($F(1,36) = 9.8$, $p < .001$, $\eta^2 = .348$).

Texts with high vocabulary were judged significantly more positively with regard to these two criteria than texts with low vocabulary.

6. Discussion

The purpose of this study was to investigate the influence of text quality and vocabulary on teachers' judgements of ESL argumentative essays to explore the role that lexical features play in teachers' analytic and holistic judgement. This research is embedded in a larger effort to identify key factors influencing assessment processes to improve classroom assessment and teacher education. Teachers carry implicit standards, or "guild knowledge" according to Sadler (1989), which are created by exchanging student work among teachers or collaborating in making assessments. Furthermore, teachers carry a history of previous judgements of text quality that varies according to their level of expertise (Sadler, 1989). Thus, teachers are strongly influenced by the range of quality that

Table 6
Variable Overall text quality.

	Low overall text quality		High overall text quality		Mean difference	d
	M	*SD*	*M*	*SD*		
Holistic scale	3.20	1.23	4.14	1.31	.94*	.74
Frame of essay	2.22	.79	2.95	.94	.73*	.84
Body of essay	2.27	.74	2.76	.89	.49*	.60
Support of arguments	2.31	.89	2.96	.87	.65*	.74
Spelling	2.45	.68	2.95	.64	.50*	.76
Grammar	2.08	.77	2.91	.80	.83*	1.06
Vocabulary	2.03	.76	2.76	.84	.73*	.91
Overall task completion	2.24	.89	2.78	.95	.54*	.59

Note. * $p < .001$.

Table 7
Variable Vocabulary.

	Low lexical richness		High lexical richness		Mean difference	d
	M	*SD*	*M*	*SD*		
Holistic scale	3.49	1.31	3.85	1.13	.36*	.29
Frame of essay	2.45	.88	2.72	.81	.27*	.32
Body of essay	2.51	.79	2.51	.94	.00	.00
Support of arguments	2.65	.76	2.62	.92	−.03	.03
Spelling	2.62	.65	2.77	.76	.15	.21
Grammar	2.34	.69	2.65	.80	.31**	.41
Vocabulary	2.14	.79	2.65	.86	.51***	.62
Overall task completion	2.42	.99	2.61	.87	.19	.20

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

exists, for example, in a set of student texts and adapt their standards in the process (Sadler, 1989). By manipulating the range of qualities of lexical features within authentic student texts, we aimed to uncover the nature of such implicit standards and how they might shift under the influence of different qualities of lexical features in student texts. Using the SIA, we implemented an experimental approach to study teacher judgements with high internal validity and linguistic qualities. The online survey instrument enabled participants to assess authentic and complex ESL argumentative essays using holistic and analytic rating scales in a realistic environment in which different aspects of text quality could be manipulated experimentally. Although there are limitations to such an approach, which we discuss below, we feel that an experimental study of teacher assessment involving independent variables such as overall text quality and lexical features is relevant since argumentative essays are an essential component of the upper-secondary ESL curriculum in Europe, and lexical features are important predictors of essay quality.

6.1. Teacher judgements of overall text quality

The first research question focused on the ability of pre-service teachers to differentiate between text with different levels of overall text quality, as measured by the NAEP framework (Board, 2010). The results indicated that the pre-service teachers in our sample were indeed able to distinguish texts with a high overall text quality from texts with a low overall text quality with regard to all criteria. This is a positive outcome for systems of teacher education in the two countries involved since the sample consisted of pre-service teachers with low to middle teaching experience. Additionally, these results provide some insight into pre-service teachers' judgements and their ability to assess the quality of upper-intermediate ESL argumentative essays on different rating scales.

6.2. Teacher judgements of manipulated texts

The second research question focused on the degree to which lexical features in ESL argumentative essays influence teachers' judgements of other analytic criteria. Essentially, this question was aimed at exploring whether a halo effect can be observed concerning the assessment of other analytic criteria. Furthermore, this study examined whether the level of lexical diversity and sophistication in students' written compositions can successfully be manipulated following the three measures of MTLD, D, and word range, calculated by the two programmes Coh-Metrix and TAALES.

With regards to a possible halo effect, the results indicate that texts with greater lexical diversity and sophistication were judged more positively with regard to the criterion *frame of essay*. This feature asked participants to judge whether a text had an effective introduction with a hook and thesis statement and an effective conclusion summarizing the main arguments. As this criterion is not directly related to lexical richness, this indicates a distorted judgement that would lead to imprecise feedback in a formative assessment situation or a faulty grading of the essay in a summative one. This, in turn, might result in students losing motivation or confidence or – if continued over a period of time and compounded with other factors – diminished opportunities of being transferred to the next educational level. Although this study identified just one such halo effect, we believe that drawing teachers' attention to such phenomena, and backing it up with empirical data, will serve to improve the quality of teacher education in this particular field.

The effect of lexical diversity and sophistication on participants' assessment of grammar indicates that pre-service teachers recognize a connection between vocabulary and grammar. Indeed, according to Nation and Webb (2011), one aspect of lexical richness encompasses grammatical accuracy of sentences. With regard to educational practices, prospective teachers should be made aware of the influence of vocabulary on the perceived grammatical accuracy. This would be conducive to fairer and more objective assessments of students who possess a large productive vocabulary yet are struggling with grammatical accuracy, for example. It might also make formative feedback more helpful and concrete.

The findings further suggest that participants assessed texts with greater lexical diversity and sophistication more positively with regard to the analytic criterion *vocabulary* and the holistic scale, indicating a successful manipulation. The tools presented in this study are not only helpful for the quantification of the relationship between writing proficiency and lexical features but also suitable for measuring the manipulation of a learner's productive vocabulary in an experimental research design. In line with previous research, lexical diversity and sophistication proved to be predictors of teachers' judgements of writing proficiency and can be considered important for the quality of written compositions (2014b, Crossley & McNamara, 2011; Crossley, Kyle et al., 2014; Kyle &

Crossley, 2016). Furthermore, the measures D, MTLD and word range facilitated the analysis and variation of lexical diversity and sophistication in the four learner texts.

With regard to the lens model, our findings indicate that vocabulary has an influence on the observed judged achievement estimates of other textual features. Thus, vocabulary is a cue that distorts the lens of teachers' judgement of *grammar* and *frame of essay*. Furthermore, we conclude that lexical sophistication and diversity are sound cues for inferring the true achievement level of the dimension of vocabulary in a student text.

7. Implications

7.1. Implications for research and practice

This study aims to make a practical impact by giving diagnostic competence a more prominent place in the curriculum of teacher education. Even though this study is limited in the sense that it cannot explain the collaborative processes in which teachers form the type of "guild knowledge" (Sadler, 1989, p. 126), we believe that implementing the SIA in teacher training seminars – as well as disseminating the results of our research – could have a positive impact on the development of teachers' diagnostic competence. Working with the SIA and exchanging their experiences with peers could help prospective teachers understand the influence of salient factors in ESL writing assessment, improve their abilities and possibly raise their interest in practising assessment. Working with the SIA could further teachers' ability to make high-quality judgements of writing by providing them with more than rubrics and criteria on which to base their assessment. Furthermore, the SIA supports them in gaining experience in making multiple judgements of texts in which categories relevant to the rubrics have been systematically manipulated. By engaging them in this activity, the SIA provides them with experience in verbalizing their reasons and discussing them with appropriate colleagues. As such, it would make explicit a type of knowledge that typically exists only implicitly inside teachers' heads as tacit knowledge (Sadler, 1989). Changing those forms of knowledge without the explicit approach provided by an instrument such as the SIA might be hard since teachers' approaches to assessment are typically linked with deeply rooted assumptions about the psychology of learning (Black & Wiliam, 1998). Often, teacher training programmes devote a limited amount of time to discussing writing assessment, even though it is a central element in teachers' diagnostic competence (Hamp-Lyons, 2016; Weigle, 2007). The difficulties pre-service teachers encounter in writing assessment point to the need for adequate training in assessing writing. To act professionally inside and outside the classroom, future ESL teachers need solid knowledge of the factors influencing teacher judgements. The findings of this study strengthen the argument that diagnostic competence should become a specific feature of ESL teacher education, particularly where the assessment of complex, many-faceted student performances is concerned (Sadler, 1989).

7.2. Limitations and directions for future research

Despite the valuable contributions of this study, some of the present findings should be interpreted with caution. The experimental research design created an artificial situation where vocabulary was deliberately manipulated to examine an effect (see Rezaei & Lovorn, 2010). As naturally occurring student texts are free of such manipulations, it is unclear to what degree our results can be generalized to a real assessment situation. However, we aimed for the texts used in the study to mimic naturally occurring ones and manipulated the levels of lexical diversity and sophistication to lie within the range found in student texts in the particular class. The relatively low ICC coefficients can be interpreted as a result of our strict experimental approach with minimal experimental variation of text quality and vocabulary. To retrieve ecologically representative ICC values, the presented texts would need to vary on all meaningful text characteristics in the same way that they would vary in a classroom setting. Future research may follow this rationale; however, it becomes increasingly more difficult to control for additional text characteristics when letting them vary freely.

We chose an experimental design with multivariate analyses of variance to investigate the effects of text variation on teacher judgements, whereas a substantial body of literature evaluates rating quality by estimating latent rater and text scores through item response analyses (Eckes, 2015; Engelhard & Wind, 2018). This approach enables the retrieval of information on the rating severity, the consistency of rating severity of raters across domains, and text quality corrected for rater severity. Under conditions of good model fit, this would allow for an estimation of the probability of a text receiving a certain rating on an analytic rubric from a specific rater. However, we believe that the multi-facet Rasch model requires either a substantially large dataset both in the number of texts and the number of raters to estimate reliable text parameters or sufficiently reliable external estimations of text parameters that can be fed into the model. In our study, we relied on a smaller dataset and did not introduce fixed text parameters beyond the simple dichotomous differentiation of overall text quality and the experimental variation of vocabulary. Thus, we confined ourselves to our original experimental approach of varying one text parameter so that we could utilize two versions of the same text to estimate the exact effects of the text variation.

Many studies stress the need for a locally developed assessment rubric following a specific purpose and aimed at a specific group of learners (Rakedzon & Baram-Tsabari, 2017; Rezaei & Lovorn, 2010). This study adapted its analytic rating scales based on existing scales, ESL learners' previous instruction and their syllabus, and two rounds of piloting and revision. A further limitation of this study is that these rating scales were developed by ESL experts at an institute of teacher education and were not tested with a large sample of teachers in the field prior to implementation. A closer analysis of the teachers' experience in assessing students' written compositions using rating scales is desirable, and we aim to do so in our future research.

Another limitation involves the selection of lexical features, which does not reflect the entire breadth and depth of vocabulary in a student's written composition. The manipulation of two components of lexical richness excludes various aspects of lexical knowledge,

such as lexical and semantic errors, and grammatical accuracy (Nation & Webb, 2011). A more comprehensive choice of lexical features may influence the analytic assessment of student texts differently; this could again be examined in future research. Additional research that explores the influence of lexical features on teacher judgements with various types of writing data is also needed to make more generalizable suggestions, for example, different writing prompts, genres or proficiency levels.

Finally, the aim of this study was to investigate the influence of lexical features on teachers' assessment of other text characteristics. This study contributes to our theoretical understanding of teacher judgements of ESL essays and suggests that our next step should be to examine other determinants of teacher judgements. For instance, this study did not discuss pre-service teachers' ratings in comparison to expert scores within the context of teacher judgement accuracy and bias. Based on the findings of this study, future studies are planned using the SIA to examine the influence of teaching experience and *pedagogical content knowledge* (Shulman, 1987) on teacher judgements, especially in comparison to expert ratings. Further studies involve text characteristics such as spelling and organization, student characteristics such as gender and migration background, as well as judgement factors such as the influence of a prompt instructing participants to be aware of possible distortion effects. We believe that these studies will improve our understanding of teacher judgements of ESL essays and provide empirical data on a key aspect of diagnostic competence. Furthermore, the SIA could be used to train teachers' diagnostic competence, particularly to raise their awareness of halo effects when evaluating written student compositions. This, in turn, would make formative assessment more precise and helpful for student development, make summative assessment fairer and more objective, and thus reduce the level of arbitrary judgements and classroom effects within the educational system as a whole.

Funding

This work was supported by the Swiss National Science Foundation [165483] and the German research foundation [Mo648/25-1].

References

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236–260.
Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaften, 4*, 469–520.
Bearman, M., Dawson, P., Boud, D., Bennett, S., Hall, M., & Molloy, E. (2016). Support for assessment practice: Developing the assessment design decisions framework. *Teaching in Higher Education, 21*(5), 545–556.
Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement, 34*(8), 607–619.
Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Erlbaum Associates: Hillsdale.
Birkel, P., & Birkel, C. (2002). How concordant are teachers' essay scorings? A replication of Rudolf Weiss' studies. *Psychologie in Erziehung und Unterricht, 49*, 219–224.
Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education Principles Policy and Practice, 5*(1), 7–74.
Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623–637.
Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on the formative and summative evaluation of student learning.* New York: McGraw-Hill.
Board, N. A. G. (2010). *Writing framework for the 2011 national assessment of education progress.* Washington, DC: US Government Printing Office.
Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.
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). .
Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6 + 1 trait writing model on grade 5 student writing achievement - final report*Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Council, o. E. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.
Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning, 21*(2/3), 170–191.
Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115–135.
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263.
Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based incides of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*, 965–981.
Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing Evaluation. *Journal of Writing Assessment, 7*(1).
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis, & M. Daller (Eds.). *Vocabulary knowledge - Human ratings and automated measures* (pp. 105–134). Amsterdam: John Benjamins Publishing Company.
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570–590.
Crusan, D. (2010). *Assessment in the second language writing classroom.* United States of America: University of Michigan.
Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43–56.
Culham, R. (2003). *6 + 1 traits of writing: The complete guide.* New York: Scholastic.
Cumming, A., Kantor, R., Baba, K., Erdosy, U. M., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(i), 67–96.
Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.
Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185.
Eckes, T. (2015). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. *Frankfurt am Main: Peter Lang.*.
Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In D. Tsagari, & J. Banerjee (Eds.). *Handbook of second language assessment* (pp. 147–164). Boston: de Gruyter.
EDK, S. K.d.k. E. (1994). *Rahmenlehrplan für die Maturitätsschulen.* Bern: EDK.

EDK, S. K.d.k. E. (1995). *Reglement über die Anerkennung von gymnasialen Maturiätsausweisen.* Bern: EDK.
Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139–155.
Engelhard, G., & Wind, S. A. (2018). *Invariant Measurement with Raters And Rating Scales: Rasch Models for Rater-Mediated Assessments.* New York: Taylor & Francis.
Federal Office for Statistics, B (2014). *Educational achievement (Bildungsabschlüsse) education and science (Bildung und Wissenschaft), vol. 15*. Neuchâtel: Office fédéral de la statistique (OFS).
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.
Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *National Council of Teachers of English, 32*(4), 365–387.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments & Computers, 36*, 193–202.
Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145.
Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*, 218–238.
Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.). *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge: Cambridge University Press.
Hamp-Lyons, L. (2016). Purposes of assessment. In D. Tsagari, & J. Banerjee (Eds.). *Handbook of second language assessment* (pp. 13–27). Boston: de Gruyter.
Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*, 122–159.
Hyland, K. (1990). A genre description of the argumentative essay. *RELC Journal,* (21), 66–78.
Hyland, K. (2008). *Second language writing.* New York: Cambridge University Press.
Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik [Handbook of pedagogical diagnostics].* Weinheim: Beltz Verlag.
Jarvis, S. (2002). Short texts, best fitting curves, and new measures of lexical diversity. *Language Testing, 19*, 57–84.
Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis, & M. Daller (Eds.). *Vocabulary knowledge - Human ratings and automated measures* (pp. 13–43). Amsterdam: John Benjamins Publishing Company.
Jarvis, S., & Daller, M. (2013). Introduction. In S. Jarvis, & M. Daller (Eds.). *Vocabulary knowledge - Human ratings and automated measures* (pp. 1–11). Amsterdam: John Benjamins Publishing Company.
Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*, 377–403.
Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaften, 18*, 279–302. https://doi.org/10.1007/s11618-015-0619-5.
Kaiser, J., Südkamp, A., & Möller, J. (2016). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871–888.
Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe - Theorie, Prozessgestaltung, Empirie [an integrative approach of teaching English writing at secondary level - theory, process development, empiricism].* Tübingen: Narr Verlag.
KMK (2012). In S. KMK (Ed.). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch / Französisch) für die Allgemeine Hochschulreife* Berlin.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275–304.
Knopp, M., Jost, J., Nachtwei, N., Becker-Mrotzek, M., Grabowski, J., et al. (2012). Teilkomponenten von Schreibkompetenz untersuchen: Bericht aus einem interdisziplinären empirischen Projekt. In H. Bayrhuber, U. Harms, B. Muszynski, B. Ralle, M. Rothgangel, & L.-H. Schön (Eds.). *Formate Fachdidaktischer Forschung - Empirische Projekte - historische Analysen - theoretische Grundlegungen* (pp. 47–65). Berlin: Waxmann.
Kronig, W. (2007). *Die systematische Zufälligkeit des Bildungserfolgs [The systematic continguency of educational achievement].* Bern: Haupt Verlag.
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.
Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.
Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Hove, UK: Language Teaching Publications.
Linnarud, M. (1986). *Lexis in composition: A performance analysis of swedish learners' written english.* Malmö: Liber Förlag.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal, 96*(ii), 190–208.
Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* London: Palgrave Macmillan.
Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal, 4*(4), 375–385.
McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the Measure of Textual, Lexical Diversity (MTLD). Unpublished doctoral dissertation.* University of Memphis.
McCarthy, P. M., & Jarvis, S. (2007). *Vocd*: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.
McCarthy, P. M., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment - An ecological validation study. In S. Jarvis, & M. Daller (Eds.). *Vocabulary knowledge - Human ratings and automated measures* (pp. 45–77). Amsterdam: John Benjamins Publishing Company.
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix.* Cambridge: Cambridge University Press.
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.
Meier, V., Trace, J., & Janssen, G. (2016). Principled rubric adoption and adaptation: One multi-method case study. In J. Banerjee, & D. Tsagari (Eds.). *Contemporary second language assessment* (pp. 165–188). London: Bloomsbury Academic.
Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* Princeton, New Jersey: Educational Testing Service.
Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston: Heinle.
Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*, 45–65.
Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*, 68–85.
Porsch, R. (2010). *Schreibkompetenzvermittlung im Englischunterricht in der Sekundarstufe I - Empirische Analysen zu Leistungen, Einstellungen, Unterrichtsmethoden und Zusammenhängen von Leistungen in der Mutter- und Fremdsprache.* Münster: Waxmann.
Porsch, R., & Köller, O. (2010). Standardbasiertes Testen von Schreibkompetenzen im Fach Englisch. In W. Bos, E. Klieme, & O. Köller (Eds.). *Schulische Lerngelegenheiten und Kompetenzentwicklung* (pp. 85–103). Münster: Waxmann.
Rakedzon, T., & Baram-Tsabari, A. (2017). To make a long story short: A rubric for assessing graduate students' academic and popular science writing skills. *Assessing Writing, 32*, 28–42.
Rauin, U., & Meier, U. (2007). Subjektive Einschätzungen des Kompetenzerwerbs in der Lehramtsausbildung. In M. Lüders, & J. Wissinger (Eds.). *Forschung zur Lehrerbildung - Kompetenzentwicklung und Programmevaluation* (pp. 103–131). Münster: Waxmann.
Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.
Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18–39.
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.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.
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.
Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology, 6*(4), 147–151.
Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerbildung, 31*(2), 154–165.
Scriven, M. (1967). *The methodology of evaluation.* Washington, DC: American Educational Research Association.
Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*(1), 27–33.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.
Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–21.
Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17–35.
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.
The British National Corpus, & v. B. X. E (2007). *From oxford university computing services on behalf of the BNC consortium.* http://www.natcorp.ox.ac.uk.
Thorndike, E. L. (1920). A constant error in psychological ratings. *The Journal of Applied Psychology, 33*, 263–271.
Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French - An exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis, & M. Daller (Eds.). *Vocabulary knowledge - Human ratings and automated measures* (pp. 79–103). Amsterdam: John Benjamins Publishing Company.
Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.
Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*, 194–209.
Weir, C. (1988). The specification, realization and validation of an English language proficiency test. In A. Hughes (Ed.). *Testing English for university study. ELT documents 127* (pp. 45–110). London: Modern English Publications in association with The British Council.
White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal, 3*(1), 3–25.
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.
Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10.
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259.
Zemach, D. E., & Stafford-Yilmaz, L. (2008). *Writers at work - the essay.* New York: Cambridge University Press.
Zhu, W. (2001). Performing argumentative writing in english: Difficulties, processes, and strategies. *TESL Canada Journal, 19*(1), 34–50.

Cristina Vögelin is a Ph.D. candidate in Educational Sciences at University of Basel. She is also working as a research associate at the School of Education, University of Applied Sciences and Arts Northwestern Switzerland. Her research interests include second language acquisition, language assessment and corpus linguistics.

Thorben Jansen is a Ph.D. candidate in Psychology at Kiel University. He is also working as a research associate at the Institute for Psychology of Learning and Instruction, Kiel University. His research interests are diagnostic competence, text assessment and ESL learning.

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.

Nils Machts is a Ph.D. candidate in Psychology at Kiel University. He is also working and lecturing as a research associate at the Institute for Psychology of Learning and Instruction, Kiel University.

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