RISTAL 2 / 2019

Research in Subject-matter Teaching and Learning

Volume 2

Citation:
Keller, S.D., Vögelin, C., Jansen, T., Machts, N. & Möller, J. (2019). Can an instructional video increase the quality of English teachers' assessment of learner essays? *RISTAL, 2*, 140–161.

DOI: https://doi.org/10.23770/pending

ISSN 2616-7697

This work is licensed under a Creative Commons Attribution 4.0 International License. (CC BY 4.0)

RISTAL. Research in Subject-matter Teaching and Learning 2 (2019), pp. 118–139

Can an instructional video increase the quality of English teachers' assessment of learner essays?

Stefan D. Keller, Cristina Vögelin, Thorben Jansen, Nils Machts & Jens Möller

Abstract

Teacher judgments of student achievement influence how effectively students learn and how teachers organize their lessons. Teachers' diagnostic competence is therefore an important field of research for subject-specific teaching and learning. This study investigates how pre-service English teachers assess complex learner essays using assessment rubrics. In particular, it explores whether instructional videos are effective in minimizing distortion effects in essay assessment by raising participants' awareness about them. English pre-service teachers (*N* = 81) in Switzerland and Germany assessed four argumentative essays of varying lexical and overall quality. Prior to the assessment task, the treatment group saw a ten-minute video about how to assess vocabulary, while the control group watched a video about general aspects of assessment. In their judgments, pre-service teachers recognized higher/lower overall quality and higher/lower lexical quality. More importantly, the results showed significant effects of the variation in lexical quality on other text characteristics, indicating halo effects. These effects were similar in both groups, suggesting that the video instruction on its own was no safeguard against distortion effects in essay assessment. Implications for teacher education and further research are discussed.

Keywords

Teaching English as a Second or Other Language (TESOL), Argumentative Essay, Assessment, Teacher education, Instructional video

1 The importance of diagnostic competence for subject-matter teaching and learning

Teachers' diagnostic competence is an important and lively field of research (Südkamp & Praetorius, 2017) which is of great importance to subject-matter teaching and learning. Teacher judgments are often the primary source of information on students' academic achievement. The ability to assess student achievement accurately is considered to be an important aspect of teachers' professional competence (Südkamp, Kaiser, & Möller, 2012; Baumert & Kunter, 2006). In the widest sense, diagnostic competence is defined as a combination of pedagogical attitude towards learners (interest in their develop-ment); hermeneutic abilities (seeing, observing and interpreting relevant information); skills in selecting diagnostic material (collecting evidence for student learning); designing and adapting tests and assessments; and applying the results of diagnosis in further learning (Edelenbos & Kubanek-German, 2004). In empirical research, a narrower defi-nition is usually chosen and diagnostic competence is described as the ability to correctly assess student and task characteristics (Schrader, 2013).

Diagnostic competence is vital not only in order to make assessment correct, fair and transparent, but also for classroom practice, for example when teachers align lessons

DOI: https://doi.org/10.23770/pending

with the perceived competences of their students (Brookhart & Chen, 2015; Hoge & Coladarci, 1989). The quality of teacher judgments has consequences for their instructional practice, for the further evaluation of students' performances, and for placement decisions – and it can influence individual students' academic careers and self-concepts (Südkamp, Kaiser, & Möller, 2012). Studies have shown that accurate teacher judgments can help to identify children who show early signs of difficulties in school (Bailey & Drummond, 2006), and that accurate information on students' academic achievement is crucial for meaningful placement decisions (Helwig, Anderson, & Tindal, 2001).

A major development in conceptualizing diagnostic competence in recent years has been a shift in the focus of attention, towards greater interest in the interactions be-tween diagnosis and classroom learning (Black & Wiliam, 1998). It became increasingly clear that teachers' judgments influence their selection of classroom activities and ma-terials; that they impact on the difficulties of the tasks selected, the choice of question-ing strategies, and the organization of student learning groups – and that they may even prompt teachers to revise their teaching methods (Südkamp, Kaiser, & Möller, 2012). Further, it has been shown that various forms of diagnosis – such as regularly assessing student progress and giving corrective feedback – directly influence student perfor-mance (Creemers, 1994). The COACTIV Study showed a significant link between Mathe-matics teachers' diagnostic competences and the learning gains of their students (An-ders, Kunter, Brunner, Krauss, & Baumert, 2010). The link between diagnostic compe-tence and classroom learning is so strong because teachers cannot teach adaptively without adequate knowledge of the achievements and learning conditions of their stu-dents (Südkamp & Praetorius, 2017). Teaching should ideally be based on the learning requirements of the students, which means that it should neither be overly demanding nor boring, and should respect learners' motivational and emotional states. This combi-nation, which is also known as formative assessment (Black & Wiliam,1998), is regarded as one of the most effective framework concepts for promoting school learning (Hattie, 2009; Visible Learning Plus, 2018), but only if the assessment is sufficiently accurate (Schütze, Souvignier & Hasselhorn, 2018).

Given the great importance of diagnostic competence for teaching and learning, one would expect a large body of subject-specific empirical research on the topic. This, how-ever, is not the case. There is a range of studies on general aspects of teacher judgment accuracy, such as the correlation between teacher judgments (grades) and student per-formance on standardized tests (for example, Feinberg & Shapiro, 2009). Such studies are not directly applicable to subject-specific assessment tasks, however, which typically involve observing and interpreting student behavior in complex performance situations or assessing the outcomes of such processes (i.e., texts, utterances, artefacts) with rel-evant criteria. According to the model by Südkamp, Kaiser, & Möller (2012), quality of assessment in any educational situation is determined by four types of characteristics: teacher characteristics (knowledge of the subject and assessment techniques, beliefs about learning); student characteristics (gender, ethnicity, level of previous knowledge); judgment characteristics (norm-referenced vs. peer-dependent judgments; domain specificity), and test characteristics (type of test used in assessment). Only a few studies, however, have examined the interplay of these characteristics for specific educational domains or topics.

DOI: https://doi.org/10.23770/pending

Where such studies have been undertaken, it appears that diagnostic competence is a situation-specific ability and cannot easily be transferred from one subject or domain to another. A study by Karing (2009) involved both subject-specific domains (arithmetic, vocabulary, text comprehension) and general domains of diagnosis (learners' subject specific interests), as well as different types of teachers (elementary school vs. Gymnasium). It found that all teachers judged subject-specific domains with greater accuracy than general ones and that diagnostic competence of elementary school teachers was generally higher than that of secondary school teachers.

A study by Klose (2014) measured how well teachers of religious education were able to assess students' religious values and domain-specific knowledge of 'theology and science'. It found that teachers systematically overestimated how much values held by individual students aligned with their own. Further, they were able to assess traditionally 'relationship-oriented' students more accurately than other types of students, suggesting a direct influence of student characteristics on judgment accuracy. By contrast, a study on English writing assessment by Jansen, Vögelin, Machts, Keller, Möller (2019) showed that teacher judgments of upper-secondary student essays were largely unaffected by student characteristics: while pre-service teachers were able to distinguish between texts of stronger and weaker overall quality, their judgments were not influenced by students' gender nor ethnic background.

Taken together, these studies suggest that assessment tasks differ significantly between subjects, and that diagnostic competence exists in varying configurations mirroring the variability of the assessment domains and tasks. The implications for subject-specific research on teaching and learning are twofold. First, there should be more studies examining subject-specific assessment tasks together with the factors influencing the quality of diagnosis pertaining to specific domains and tasks. Second, there should be more studies evaluating programs to foster diagnostic competence in those domains and tasks. Although some preliminary approaches to fostering diagnostic competence in different contexts have been published (e.g., Förster & Souvignier (2015) for reading processes; Klug, Gerich, Bruder, & Schmitz, (2012) for teacher diaries), a systematic overview of programs supporting diagnostic competence has hitherto been lacking – and the same goes for empirical research on the effectiveness of such programs (Südkamp & Praetorius, 2017).

This study hopes to make a small but significant contribution in both areas. First, it focuses on a key assessment task of (foreign) language writing, investigating how the quality of learners' vocabulary influences the judgment of other, unrelated text characteristics in criteria-based assessment. Second, it tests whether a short, video-based instruction is effective at improving teachers' diagnostic competence in this specific assessment task.

DOI: https://doi.org/10.23770/pending

2 Diagnostic competence in assessing (foreign language) learner texts

2.1 Importance of foreign language writing and assessment

For the last two decades, foreign language education has received increasing attention in the European Union, where citizens are expected to be proficient in two foreign languages in addition to their first when they complete secondary education (European Commission, 2008). Thus, there is a need both for high-quality foreign language instruction in schools, and for close monitoring of its effectiveness. Teachers' diagnostic competence is especially relevant in the field of English writing because this is a key competence for higher education in a globalized world in general, and for tertiary education in particular (Keller, 2013). In a general sense, foreign languages teachers' diagnostic competence could be defined as the ability to interpret students' foreign language growth, to deal with assessment material skillfully and to provide students with appropriate help in response to this diagnosis (Edelenbos & Kubanek-German, 2004). Diagnostic competence includes the use of formal assessment tools such as tests as well as informal assessment tools such as observation and feedback during lessons. When assessing student writing in particular, teachers require a solid grasp of text quality and should be able to assign fair judgments based upon transparent criteria (Vögelin, Jansen, Keller, Machts, & Möller, 2019; Hyland, 2008; Weigle 2002). Distinguishing between different text characteristics, or combining them into an accurate overall assessment of text quality, requires specialized knowledge of genre, linguistic structures and relevant assessment criteria (Cooksey, Freebody, & Wyatt-Smith, 2007). In the context of this study, therefore, we refer to the terms "diagnosis" and "diagnostic competence" in the sense of teachers rating student texts with specific criteria.

Teachers often manage this complexity by employing assessment rubrics as guides to classifying essays into clearly defined categories intended to indicate levels of writing (Goodrich Andrade, 2005). An assessment rubric is the instrument which provides criteria for determining the quality of the written language sample (Shohamy, Gordon, & Kraemer, 1992). "Analytic" rubrics require teachers to assign several trait scores, each of which depicts the quality of a particular aspect of the writing. "Holistic" rubrics, by contrast, require teachers to consider all relevant aspects of an essay in arriving at a single score expressing the overall quality of the writing (Wolfe & Jiao, 2014). The use of assessment rubrics has been shown to contribute to the fairness and objectivity of assessment at different levels of writing proficiency (Goodrich Andrade, 2005). By the same token, teachers are prone to assessment errors while applying analytic rubrics, for example when one characteristic of text quality unduly influences the assessment of another ("halo effect"; Rezaei & Lovorn, 2010).

Several studies have investigated how specific text characteristics influence teachers' assessment of student texts (see Vögelin et al., 2019, for an overview of relevant research). For example, both pre-service and experienced teachers tend to assign lower grades to essays containing mechanical errors (Cumming, Kantor, & Powers, 2002; Birkel & Birkel, 2002). 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, e.g., understanding and

DOI: https://doi.org/10.23770/pending

synthesis of argument. Such distortion effects are particularly likely to occur in second language writing because such texts contain more errors than first language texts (Cumming, Kantor, & Powers, 2002). To compound the problem, some teachers who are non-native English speakers tend to attend more extensively to formal language issues in their assessment of student essays, in comparison to other aspects, such as organization or structure (Eckes, 2008). Jansen, Vögelin, Machts, Keller, & Möller (2018), however, showed that student teachers who received a prompt cautioning them against a possible judgment error caused by spelling judged learner essays in a significantly less biased way.

A key research issue, therefore, is identifying how particular textual features can distort teachers' text assessment, and finding ways of inoculating teachers against those effects. In the next section, we focus specifically on the influence of lexical quality on essay assessment as this aspect is at the center of the empirical study presented here.

2.2 Influence of lexical quality on essay assessment

The quality of writing assessment is moderated by (i.e., varies as a function of) text characteristics such as spelling, vocabulary or organizational quality. These aspects of textual quality are the most immediate or "proximal" influences on the outcome of writing assessment (Eckes, 2005). The ability to recognize and use a wide range of lexical items and idiomatic expressions is a key feature of English proficiency (Lewis, 1997), and it is particularly relevant for argumentative writing (Keller, 2013). For this reason, lexical quality is an important influence on how teachers perceive the quality of learner essays (see Vögelin et al. (2019) for an extensive literature review). Lexical quality can be analyzed as *lexical diversity* (range and variety of vocabulary), and *lexical sophistication* (percentage of advanced or less frequent words; Nation & Webb, 2011). Both are important indicators of foreign language academic achievement (Daller, Milton, & Treffers-Daller, 2007). For example, Guo, Crossley and McNamara (2013) showed that lexical sophistication, among other characteristics, significantly predicted the assessment of L2 writing proficiency in the writing tasks of the *Test of English as a Foreign Language* (TOEFL).

Vögelin et al. (2019) showed that English pre-service teachers can distinguish higher and lower lexical quality in authentic student texts but that this aspect can have significant halo effects on other, unrelated textual aspects. Using analytic and holistic rating scales, participants assessed four essays of different proficiency levels in which the levels of lexical diversity and sophistication had been experimentally varied. The results suggested that texts with greater lexical quality were assessed more positively concerning their overall quality. However, the manipulations of vocabulary quality also 'spilled over' on the assessment of grammar and frame of essay, two separate categories on the assessment rubric which were judged more positively when lexical quality was higher. As this suggests a halo effect, we conducted the present study to investigate whether an instructional video might help teacher trainees to avoid such distortions of assessment.

2.3 Teachers' professional knowledge

There are various studies suggesting that qualification and professional knowledge are characteristics which help teachers in general to make more accurate judgements: teachers who have spent several years in the profession have more experience and thus more practice in evaluating student performance (Blömeke, Gustafsson, & Shavelson,

DOI: https://doi.org/10.23770/pending

2015; Glogger-Frey, Herppich, & Seidel, 2018). However, teaching experience does not automatically transfer into high diagnostic competence or accurate assessment. For example, Keller & Möller (2019) showed that experienced English teachers at upper secondary level in Switzerland and Germany consistently underestimated students' writing competences (i.e., judged their texts too harshly) in comparison to benchmarks scores. By contrast, teacher trainees were more accurate in their assessment of the same texts, possibly because they were not yet subject to an 'expert blind spot'.

Studies such as this one underline the importance of specific teacher training as a method of fostering diagnostic competence. In first language text assessment, Chamberlain & Taylor (2011) showed that both on-line and face-to-face training measures proved effective in moving assessments by experienced raters closer to those of a "principal examiner". Positive effects of on-line assessment training also emerged in a study involving pre-service teachers by Dempsey et al. (2009). Participants assessed writing samples via a web-based "critical thinking tool". The study involved both intermediate and upper-level teacher trainees with differing background knowledge of text assessment. They received scaffolded practice in assessing multiple student papers and justified their assessments using analytic criteria. Results showed an improvement in participants' ability to match experts' ratings of writing quality, irrespective of previous knowledge. However, the study did not focus on halo effects, nor did it have a control group which did not have access to the training.

The available literature suggests that improving assessment criteria, training teachers/examiners, fostering a community of practice and providing exemplary material (models) may all help to improve assessment quality (Meadows & Billington, 2010). Yet while such training can be effective, there appears to be no simple input-output relation between training and assessment quality. For example, Royal-Dawson and Baird (2009) compared ratings by specially trained assessors to those by untrained teachers and students for two English text assignments in the General Certificate of Secondary Education (GCSE) test. Comparison of assessments from these groups with those of the most experienced assessor showed no significant differences. Meadows and Billington (2010) replicated the study with two other English assignments from the GCSE and a larger sample and again found no differences between the groups. In sum, it is unclear to what extent professional knowledge leads to higher quality assessments, and there is a dearth of studies examining training measures for teachers rather than professional raters in high-stakes contexts.

2.4 Instructional videos in teacher training

There is consensus among researchers and teacher educators that videos can be valuable tools for teacher education (Fadde & Sullivan, 2013; Darling-Hammond, 2006). Videos can convey the complexity and subtlety of classroom teaching as occurring in real time with a richness and immediacy that written descriptions cannot achieve. For example, classroom videos can allow teacher trainees to be immersed in a classroom without the pressure of having to interact or make real-life evaluations of student productions. In addition, instructional videos can serve as a necessary bridge between theory and practice: they can be used to illustrate particular theories of teaching and learning and create opportunities for applying these theories in practice (see Blomberg et al., 2013 for an overview of this research).

DOI: https://doi.org/10.23770/pending

There are a number of on-line instructional videos aiming to educate teachers about writing assessment. The British Council produced a video showing teachers how to design a writing test and assess the outcome[1], and so did the agency running the National Assessment of Educational Progress (NAEP) in the USA[2]. If used in the context of a training study, such s take on the function of a "non-interactive coach" (Dempsey at al., 2009): they provide teachers with new knowledge or support the re-organization of existing knowledge about a specific procedure of writing assessment. How they apply that knowledge, however, must be evaluated separately for each new context.

2.5 Purpose of the present study

The purpose of this study is to explore the use and effectiveness of video-based instruction in reducing halo effects between different criteria in the assessment of English argumentative essays. First, we want to replicate a study on the effects of overall text quality and lexical quality in which we showed that both aspects influenced other text characteristics such as grammar or essay organization (Vögelin et al., 2019). Second, we aim to explore whether such halo effects can be mitigated with a specific, video-based instruction on lexical quality. In particular, our investigation tests three hypotheses:

1. Pre-service teachers are able to distinguish higher overall text quality and higher lexical quality from respective lower qualities;
2. The level of lexical quality in ESL argumentative essays affects pre-service teachers' holistic and analytical judgement of these texts, suggesting halo effects;
3. Halo effects are smaller in the experimental group (video with vocabulary topic) than in the control group (video with general assessment topic).

All of these hypotheses were based on our previous research on the effects of linguistic text features on teacher ratings. In particular, hypothesis 3 was informed by an experimental study which tested the influence of spelling on other aspects of student essays in two groups (Jansen et al., 2018). Before the assessment, the participants in the experimental group were given a verbal prompt that instructed them to pay attention to a possible influence of spelling errors on their assessment of other text characteristics. The prompt significantly reduced the halo effect of spelling on the scales 'support of arguments', 'vocabulary', and 'overall task completion' in comparison to the control group. This present study tested whether the same positive effect could be achieved by a video. Because 'quality of vocabulary' is a more difficult concept than 'spelling', we replaced the prompt with an instructional video containing more detailed input.

3 Method

To test these hypotheses, we used an experimental 2x2x2 design with three variables: *overall text quality* (high vs. low), *lexical quality* (high vs. low) and *group* (treatment vs. control). The first two variables served as within-subject variables; group served as a between-subject variable. The mixed design was analyzed with multivariate repeated-

[1] https://www.britishcouncil.org/exam/aptis/research/projects/assessment-literacy/writing

[2] https://www.youtube.com/watch?v=NEhS2x2pvB4

DOI: https://doi.org/10.23770/pending

measures analyses of variance. The within-subject variables were identical to the study by Vögelin et al. (2019) and are described only as a summary here. The between-subject variable introduced in this study was "video with a vocabulary topic" (treatment) vs. "video with a general assessment topic" (control).

Overall text quality was operationalized in the following way: Experts from the School of Teacher Education (Basel) evaluated a total of 15 student compositions 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, while stronger texts showed levels of work that surpassed the learning goals (Vögelin et al., 2019)

Lexical quality was experimentally manipulated as "high" and "low" in the four texts. The manipulations concerned "lexical sophistication" and "lexical diversity", both of which were systematically lowered and increased in each text. To lower the level of lexical *sophistication*, less sophisticated words were substituted for more sophisticated words using frequency word lists. To lower the level of lexical *diversity*, word repetitions were inserted and used to replace more diverse vocabulary. Both characteristics of lexical quality were varied to a degree that matched the highest and lowest values nat-urally occurring in the original sample of learner texts. Results of these manipulations were checked by automated tools of lexical analysis such as of MTLD, D, Coh-Metrix and TAALES (see Vögelin at al. (2019) for details).

The between-subjects variable "group" was operationalized by showing one of two videos. In the group "video with vocabulary topic" (treatment), the video focused specifically on different aspects of lexical quality as well as possible halo effects on grammar and essay organization - effects which had been revealed as salient in an earlier study (Vögelin et al., 2019). In the group "general assessment topic" (control), the video covered general aspects of assessing writing from a standard introductory textbook (Weigle, 2002). While not irrelevant to the task, it contained no direct information about lexical sophistication and diversity nor any hint on how to distinguish these from other text characteristics. Both videos were 10 minutes long and used voice-over with visual aids. The treatment group video was piloted with a group of N = 17 students enrolled in a seminar on English language proficiency at the School of Education in Basel. Students assessed the comprehensibility, usefulness, and construct coverage of the video (evaluation ques-tionnaire, see Appendix B). The control group video was evaluated by the research team and revised accordingly.

3.1 Participants

N = 81 students of higher education in English teacher education seminars at universities in Germany and Switzerland participated in the study. The age of participants ranged from 19 to 48 years with a mean of 24.73 years (SD = 4.87). Participants were not equally divided by gender (65.4% female). The majority of participants (88.9%) had (Swiss) German as their mother tongue and described their English proficiency equivalent to level C1 (48.1%) according to the *Common European Framework for Reference* (CEFR). The remaining participants reported that their English proficiency was equivalent to C2 (39.5%), B2 (2.5%) or had English as their mother tongue (2.5%). Further, 7.4% participants stated that they did not know the level of their English proficiency. On average,

DOI: https://doi.org/10.23770/pending

participants had already completed 5.07 semesters (*SD* = 4.52) at university and reported little to medium teaching experience (*M* = 1.04 years; *SD* = 2.99). Because all of the participants were students of teacher education, it is reasonable to assume that they had attended courses on language assessment, possibly even analytic text rating. As our study was conducted in two countries at different universities and colleges with differing curricula, their level of expertise in this particular area was probably quite heterogenous.

3.2 Procedure

During the study, participants were asked to assess four authentic learner argumentative essays (11th grade), two of higher and two of lower overall quality in which the quality of vocabulary had been experimentally varied. The texts came from a teaching unit conducted in a Gymnasium in Basel (CH) and the prompt was as follows: "Do you agree or disagree with the following statement? As humans are becoming more dependent on technology, they are gradually losing their independence". Participants assessed these texts in a digital instrument called *Student Inventory ASSET* (SIA; Vögelin et al., 2019; **Jansen et al., 2019**). In this computer-based tool, participants read learner texts on the left-hand side and see rating scales displayed on the right-hand side of the screen.

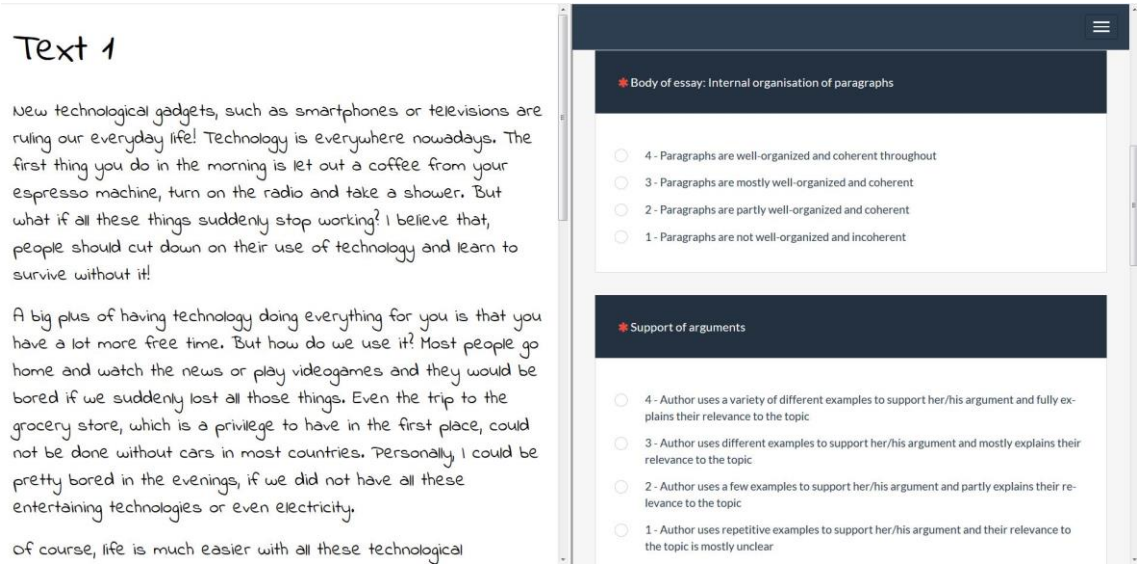

Fig. 1. Screenshot from Student Inventory ASSET (SIA)

As a first step, participants received background information on the school context in which the texts had originated: learners' age, proficiency level and subject-related previous knowledge and information on the essay prompt. Next, they read the rating scales with their different levels and descriptors. In order to familiarize themselves with the scales, participants were asked to assess a training text. They were then split randomly into a treatment group which watched a video on lexical quality and a control group which watched a video on assessment in general. The content of each video is summarized in Table 1.

DOI: https://doi.org/10.23770/pending

Table 1
Content of instructional videos used in treatment and control group

Treatment video: **Influence of lexical characteristics**	**Control video:** **Writing assessment in general**
- vocabulary as a crucial characteristic in essay quality; - two key concepts of lexical quality: lexical sophistication and lexical diversity; - concrete examples of lexical characteristics in a learner text (taken from the training text which participants assessed); - helpful tips on how to distinguish vocabulary from: • spelling • grammar • essay structure.	- quality of different rating processes; - two main types of assessing writing: holistic and analytic scoring with advantages and disadvantages; - rating scales with regard to their usefulness in terms of: • reliability • validity • practicality • impact • authenticity; - applying assessment rubrics in the classroom.

Content in the videos was presented in the form of a "whiteboard animation", as Figure 2 illustrates:

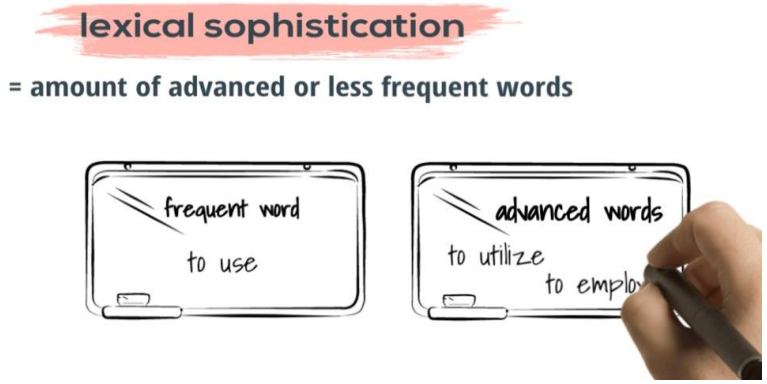

Fig. 2. Screenshot from video with vocabulary quality topic (treatment group)

After having watched the video, participants first read and then assessed the four texts in a randomized order using the holistic and analytic rating scales (see Appendix A). Each participant assessed four texts of different overall quality and different levels of lexical quality. For each text, participants were asked to assess on a holistic and an analytic scale. For the holistic rating, the well-validated NAEP rating scale with six levels using age- and grade-appropriate writing criteria was used (Board, 2010). The analytic rubric was designed for this study by adapting the *6+1 trait model* by Culham (2003). This rubric contained 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 (see Vögelin et al., 2019 for details).

DOI: https://doi.org/10.23770/pending

After they had assessed all four texts, participants filled in a background questionnaire regarding their level of education and language proficiency. Also, they answered questions about the video they had watched as a manipulation check (see Appendix C).

Results showed that 91% of participants in the treatment group answered three key questions relating to the video correctly, suggesting that they had properly absorbed the relevant information.

4 Results

To test all three hypotheses, we used a multivariate repeated-measures analyses of variance. Further, we conducted univariate post-hoc tests for all scales showing significant multivariate effects. Regarding hypothesis 1, the multivariate analysis of variance showed a significant main effect for the variable text quality ($F(8, 72) = 40.57$, $p < .001$, $\eta^2 = .82$). The univariate tests for the text quality show that texts of higher overall quality were judged more positively with regard to the holistic scale ($F(1, 79) = 132.24$, $p < .001$, $\eta^2 = .63$). Texts of higher overall quality were also assessed more positively on all analytic scales than texts with lower quality: frame of essay ($F(1, 79) = 140.54$, $p < .001$, $\eta^2 = .64$), body of essay ($F(1, 79) = 76.27$, $p < .001$, $\eta^2 = .49$), support of arguments ($F(1, 79) = 42.99$, $p < .001$, $\eta^2 = .35$), spelling ($F(1, 79) = 241.76$, $p < .001$, $\eta^2 = .75$), grammar ($F(1, 79) = 202.95$, $p < .001$, $\eta^2 = .72$), vocabulary ($F(1, 79) = 80.65$, $p < .001$, $\eta^2 = .72$), and task completion ($F(1, 79) = 80.65$, $p < .001$, $\eta^2 = .51$) (see Table 2). These results indicate that the participants differentiated between texts of high and of low overall quality on all assess-ment scales, supporting hypothesis 1.

Table 2

Influence of overall text quality on assessment on holistic and seven analytic scales

	Low text quality		High text quality		Mean difference	*d*
	M	*SD*	*M*	*SD*		
Holistic scale	**2.86**	**0.65**	**4.06**	**0.63**	**1.20**	**1.88**
Frame of essay	**2.23**	**0.51**	**3.23**	**0.55**	**1.00**	**1.89**
Body of essay	**2.54**	**0.62**	**3.26**	**0.53**	**0.72**	**1.25**
Support of arguments	**2.61**	**0.59**	**3.18**	**0.54**	**0.57**	**1.01**
Spelling	**2.31**	**0.57**	**3.40**	**0.47**	**1.09**	**2.09**
Grammar	**2.20**	**0.50**	**3.20**	**0.50**	**1.00**	**2.00**
Vocabulary	**2.40**	**0.57**	**3.07**	**0.57**	**0.67**	**1.18**
Task completion	**2.40**	**0.47**	**3.23**	**0.56**	**0.83**	**1.61**

Note. Bold analytic rating scales indicate a significant univariate difference

DOI: https://doi.org/10.23770/pending

Regarding hypothesis 2, the multivariate analysis of variance showed a significant main effect of the variable lexical quality ($F(8, 72) = 9.38$ $p < .001$, $\eta^2 = .51$). The univariate tests for the lexical quality showed that texts with high lexical quality were judged more positively with regard to the holistic scale ($F(1, 79) = 21.99$, $p < .001$, $\eta^2 = .22$) and all analytic scales than texts with low lexical quality: frame of essay ($F(1, 79) = 8.09$, $p < .01$, $\eta^2 = .09$), body of essay ($F(1, 79) = 6.59$, $p < .05$, $\eta^2 = .08$), support of arguments ($F(1, 79) = 7.02$, $p < .05$, $\eta^2 = .08$), spelling ($F(1, 79) = 9.04$, $p < .01$, $\eta^2 = .10$), grammar ($F(1, 79) = 32.60$, $p < .001$, $\eta^2 = .29$), vocabulary ($F(1, 79) = 37.11$, $p < .001$, $\eta^2 = .48$), and task completion ($F(1, 79) = 21.65$, $p < .001$, $\eta^2 = .22$; see Table 3). The results support hypothesis 2, indicating halo effects of lexical quality on all scales.

Table 3

Influence of lexical quality on assessment on holistic and seven analytic scales

	Low lexical quality		High lexical quality		Mean difference	d
	M	*SD*	*M*	*SD*		
Holistic scale	**3.23**	**0.57**	**3.69**	**0.66**	**0.46**	**0.75**
Frame of essay	**2.62**	**0.54**	**2.85**	**0.48**	**0.23**	**0.45**
Body of essay	**2.81**	**0.53**	**2.99**	**0.56**	**0.18**	**0.33**
Support of arguments	**2.80**	**0.56**	**2.99**	**0.50**	**0.19**	**0.36**
Spelling	**2.73**	**0.56**	**2.98**	**0.54**	**0.25**	**0.45**
Grammar	**2.47**	**0.58**	**2.93**	**0.49**	**0.46**	**0.86**
Vocabulary	**2.40**	**0.60**	**3.08**	**0.55**	**0.68**	**1.18**
Task completion	**2.64**	**0.54**	**2.99**	**0.53**	**0.35**	**0.65**

Note. Bold analytic rating scales indicate a significant univariate difference

Regarding hypothesis 3, the multivariate effect shows no significant interaction effect between lexical quality and group ($F(8, 72) = 1.05$, *ns*). This result does not support hypothesis 3 as the video with the vocabulary topic did not reduce halo effects of lexical quality in comparison to the one with the general assessment topic (see Table 4).

DOI: https://doi.org/10.23770/pending

Table 4

Interaction of lexical quality and group membership (vocabulary video vs. general assessment video)

	Vocabulary video (treatment)					General assessment video (control)				
	Low lexical richness		High lexical richness			Low lexical richness		High lexical richness		
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Holistic scale	3.31	0.55	3.73	0.66	0.69	3.17	0.58	3.64	0.67	0.75
Frame of essay	2.73	0.56	2.86	0.50	0.25	2.52	0.51	2.83	0.46	0.64
Body of essay	2.81	0.51	2.91	0.58	0.18	2.81	0.55	3.07	0.54	0.48
Support of arguments	2.78	0.55	3.05	0.54	0.50	2.81	0.58	2.94	0.47	0.25
Spelling	2.79	0.44	2.99	0.56	0.40	2.68	0.65	2.96	0.53	0.47
Grammar	2.55	0.55	2.99	0.48	0.85	2.39	0.60	2.88	0.50	0.89
Vocabulary	2.27	0.65	3.03	0.63	1.19	2.51	0.55	3.13	0.47	1.21
Task completion	2.67	0.59	3.04	0.54	0.65	2.61	0.49	2.94	0.52	0.65

Besides the results regarding the three hypotheses, the analyses of variance showed a significant effect of the variable "group" ($F(8, 72) = 2.26$, $p < .05$, $\eta^2 = .20$). The univariate test for the variable group showed no significant effects for the holistic scale ($F(1, 79) = 1.38$, *ns*), nor for one of the analytic scales frame of essay ($F(1, 79) = 1.98$, *ns*), body of essay ($F(1, 79) = 0.69$, *ns*), support of arguments ($F(1, 79) = 0.20$, *ns*), Spelling ($F(1, 79) = 0.55$, *ns*), grammar ($F(1, 79) = 2.29$, *ns*), vocabulary ($F(1, 79) = 3.07$, *ns*), and task completion ($F(1, 79) = 0.73$, *ns*). With this sample size, the effect of the group seems to be too small to detect with univariate tests.

Analyses also showed an interaction effect between the variables overall text quality and lexical quality ($F(8, 72) = 4.26$ $p < .001$, $\eta^2 = .32$). The univariate tests for the interaction effect between text quality and lexical quality showed effects on the holistic scale ($F(1, 79) = 12.38$, $p < .01$, $\eta^2 = .14$) and the analytic criteria frame of essay ($F(1, 79) = 17.42$, $p < .001$, $\eta^2 = .18$), support of arguments ($F(1, 79) = 4.05$, $p < .05$, $\eta^2 = .05$), grammar ($F(1, 79) = 12.66$, $p < .01$, $\eta^2 = .14$), vocabulary ($F(1,79) = 17.33$, $p < .001$, $\eta^2 = .18$), and task completion ($F(1, 79) = 6.39$, $p < .05$, $\eta^2 = .08$). The univariate analyses showed no effect for the scales body of essay ($F(1, 79) = 0.89$, *ns*) and spelling ($F(1, 79) = 0.22$, *ns*). This result indicates that the halo effects of lexical quality were larger on most scales when judging texts with high than with low overall text quality. For

DOI: https://doi.org/10.23770/pending

the means and standard deviations for overall text quality and lexical quality, see Table 5.

Table 5

Interaction of overall text quality and lexical quality

	Low text quality					High text quality				
	Low lexical quality		High lexical quality			Low lexical quality		High lexical quality		
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Holistic scale	**2.79**	**0.82**	**2.94**	**0.86**	**0.18**	**3.68**	**0.91**	**4.43**	**0.87**	**0.85**
Frame of essay	**2.28**	**0.73**	**2.19**	**0.69**	**0.14**	**2.96**	**0.81**	**3.51**	**0.65**	**0.74**
Body of essay	2.48	0.81	2.60	0.77	0.16	3.14	0.68	3.38	0.64	0.37
Support of arguments	**2.59**	**0.77**	**2.63**	**0.71**	**0.05**	**3.00**	**0.84**	**3.36**	**0.68**	**0.47**
Spelling	2.17	0.72	2.44	0.82	0.35	3.30	0.71	3.51	0.53	0.34
Grammar	**2.10**	**0.72**	**2.31**	**0.74**	**0.29**	**2.84**	**0.78**	**3.56**	**0.52**	**1.10**
Vocabulary	**2.19**	**0.67**	**2.62**	**0.77**	**0.60**	**2.60**	**0.83**	**3.54**	**0.61**	**1.30**
Task completion	**2.30**	**0.62**	**2.49**	**0.63**	**0.31**	**2.98**	**0.79**	**3.48**	**0.65**	**0.70**

Note. Bold analytic rating scales indicate a significant univariate difference

No significant interaction effect was found between lexical quality and group ($F(8, 72) = 1.05$, *ns*), text quality and group ($F(8, 72) = 1.39$, *ns*), nor between lexical quality, text quality and group ($F(8, 72) = 0.75$, *ns*).

5 Discussion

The aim of this study was to shed new light on an important assessment task of foreign language learning: criteria-based evaluation of learner texts. It analyzed the influence of overall text quality and quality of vocabulary on other text characteristics of English argumentative essays at upper secondary level. An earlier study (Vögelin et al., 2019) had shown that quality of vocabulary spilled over into the assessment of other characteristics such as grammar and text organization, suggesting halo effects. This present study replicated the earlier one and explored whether halo effects might be mitigated if preservice English teachers watched an instructional video on the effects of vocabulary quality in essay assessment.

DOI: https://doi.org/10.23770/pending

Regarding halo effects, the results of this study are in line with the earlier one: We found that overall text quality affected the assessment on the holistic and on all analytic scales, suggesting that participants were able to distinguish reliably between stronger and weaker texts. In addition, the manipulation of vocabulary quality affected all other aspects on the analytic assessment rubric, with grammar, frame of essay and task completion getting particularly strong effects. This news is not altogether bad as the manipulation of vocabulary as a text characteristic had the largest effect on the *assessment* of vocabulary, suggesting that participants did distinguish vocabulary from other text characteristics. Furthermore, the halo effects of vocabulary on other text characteristics are rather small, i.e., the manipulation did not lead to 'catastrophic' mistakes of assessment. The consistent effects of vocabulary on independent aspects such as task completion (i.e., adequately answering the essay question) or frame of essay (i.e., introduction and conclusion), however, suggest that it does spill over onto separate characteristics and distorts how these are perceived. These results prove how difficult it is for pre-service teachers to assess long and complex learner texts on several distinct criteria in a limited period of time.

Regarding the between-subject variable "group", there were no significant effects in our study: the halo effects emanating from vocabulary quality remained largely the same whether participants had watched a video targeting that very phenomenon or one that focused on general aspects of assessment only. This result underlines the need for further research on effective training tools in specific assessment situations. While instructional videos appear a promising perspective for (foreign) language teachers, it seems that they alone are insufficient to make a significant change in terms of reducing halo effects – which are quite small – when assessing learner texts. Used without further scaffolding, the videos made no significant impact in our study. We should bear in mind, however, that these videos were only 10 minutes long and there was no opportunity for participants to discuss them or practice the application of the new knowledge. In assessment studies where on-line training has proven effective, videos were used over a longer period of time and there was interaction with several partners, such as other teachers or experts (Fadde & Sullivan, 2013; Dempsey et al., 2009). Teacher training studies in other subjects, such as mathematics, used interventions which lasted for several hours, sometimes several days (Blomberg et al., 2013). Our intervention, one could surmise, may have been too short and insufficiently integrated within a "community of interpreters" (Moss, 1994) to have a significant effect.

Our study was limited by its experimental design, as there is no manipulation of vocabulary in the real world of foreign language assessment (Rezaei & Lovorn, 2010). Further, the study did not measure participants' previous knowledge of text assessment, nor did it take issues of rater reliability into account, e.g. different degrees of rater severity or leniency.

Further studies should explore subject-specific aspects of essay assessment in more 'natural' contexts. This could mean examining the interaction of vocabulary and other text characteristics in a large corpus of learner texts without experimental manipulation. In addition, further studies should explore the potential of instructional videos outside of experimental research designs. This could be achieved by incorporating them in seminars of teacher education where participants discuss the videos with peers and experts

DOI: https://doi.org/10.23770/pending

as well as practice the application of new knowledge. In such studies, one could also compare and contrast first- and foreign-language writing assessment.

On a more general level, the concept of diagnostic competence could help to build a bridge between psychometric concepts such as 'teacher judgment accuracy' and pedagogical concepts such as 'student-centered learning' (Klose, 2014). Assessment in education should be as fair and as objective as possible, ensuring correct diagnosis of learners' strengths and weaknesses, contributing to efficient learning and forming the basis for suitable placement decisions. By the same token, developing teachers' diagnostic competence is a way of sharing expert knowledge and helping them to become more professional practitioners. In trying to understand diagnostic competence in domain-specific configurations, researchers first need to account for teachers as diagnosticians in subject-specific contexts – their expertise and knowledge, their training and access to other resources, and their behavior in synthesizing various types of evidence at the decision-making stage (Alderson, Brunfaut, & Harding, 2014). This type of understanding is the basis for developing teacher training measures which will ultimately benefit the learners, both in how effectively they acquire competences in a specific subject, and how accurately the outcomes of their learning are assessed.

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNF) under Grant 165483, and the German Research Foundation (DFG) under Grant 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.

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostic competences of mathamatics teachers and the performance of their learners. *Psychologie in Erziehung und Unterricht*, 3, 175-193. doi:10.2378/peu2010.art13d

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment*, 11, 149-178. doi:10.1207/s15326977ea1103&4_2

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Re: Professional competence of teachers]. *Zeitschrift für Erziehungswissenschaften, 4*, 469-520.

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 Educa-tion: Principles, Policy & Practice*, 5:1, 7-74. doi:10.1080/0969595980050102

Blomberg, G., Sherin, M., Renkl, A., Glogger, I., & Seidel, T. (2013). Understanding video as a tool for teacher education: Investigating instructional strategies to promote reflection. *Instructional Science*, *42*, 443-463.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*, *223*, 3–13. https://doi.org/10.1027/2151-2604/a000194

Board, N. (2010). *Writing framework for the 2011 national assessment of education progress*. Washington: US Government Printing Office.

Brookhart, S., & Chen, F. (2015) The quality and effectiveness of descriptive rubrics. *Educational Review, 67/*3, 343-368. doi:10.1080/00131911.2014.929565

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

Charney, D. (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. *Research in the Teaching of English*, 18/1, 65-81.

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

Creemers, B. (1994). *The effective classroom.* London: Cassell.

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

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*, 67-96.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: CUP.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*, 120-138. doi:10.1177/0022487105

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing 2008 25* (2), 155-185. doi:10.1177/0265532207086780

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A multi-faceted Rasch analysis. *Language Assessment Quarterly, 2/*3, 197-221. doi:10.1207/s15434311laq0203_2

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: the concept of 'diagnostic competence'. *Language Testing,* 21 (3), 259-283.

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

Fadde, P., & Sullivan, P. (2013). Using interactive video to develop preservice teachers' classroom awareness. *Contemporary Issues in Technology and Teacher Education*, *13*/2, 156-174.

Feinberg, A., & Shapiro, E. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research*, 102, 453-462. doi:10.3200/JOER.102.6.453-462

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review*, 44, 60-75.

Glogger-Frey, I., Herppich, S., & Seidel, T. (2018). Linking teachers' professional knowledge and teachers' actions: Judgment processes, judgments and training, *Teaching and Teacher Education, 76*, 176-180. https://doi.org/10.1016/j.tate.2018.08.005

Goodrich Andrade, H. (2005). Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching, 53*/1, 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Guo, L., Crossley, S., & McNamara, D. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing 18/*3, 218-238. doi:10.1016/j.asw.2013.05.00

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student gender on teachers' perceptions of mathematics achievement. *Journal of Educational Research, 95,* 93-102. doi:10.1080/00220670109596577

Hoge, R., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. Review of Educational Research, 59, 297-313. doi:10.2307/1170184

Hyland, K. (2008). *Second language writing*. New York: CUP.

Jansen, T., Vögelin, C., Machts, N., Keller, S., & Möller, J. (2019). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen [The Student Inventory ASSET for judging students' performances in English: Three experimental studies on effect of text quality and student names]. *Psychologie in Erziehung und Unterricht* 66, *303-315.* doi:10.2378/peu2019.art21d

Jansen, T., Vögelin, C., Machts, N., Keller, S., & Möller, J. (2018). The influence of spelling and prompting on teacher judgments of English essays. Paper presented at the Anual Meeting of American Educational Research Association (AERA). New York City, 16. April 2018.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of elementary and secondary school teachers in the domains of competence and interests]. *Zeitschrift für Pädagogische Psychologie*, 23, 197-209. doi:10.1024/1010-0652.23.34.197

Keller, S. & Möller, J. (2019). Das Schülerinventar zur Beurteilung von Schülertexten. In: T. Rieke-Baulecke (Hrsg.). *Schulmanagement Handbuch* 169, 55-65.

Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe. Theorie, Prozessgestaltung, Empirie*. [Integrated writing instruction for secondary level – theory, processes and empirical evaluation]. Tübingen: Narr.

Klose, B. (2014). *Diagnostische Wahrnehmungskompetenzen von ReligionslehrerInnen* [Diagnostic competences of teachers of religious education]. Stuttgart: Kohlhammer.

Klug, J., Gerich, M., Bruder, S., & Schmitz, B. (2012). Ein Tagebuch für Hauptschullehrkräfte zur Unterstützung der Reflexionsprozesse beim Diagnostizieren [A diary for teachers to support reflection on diagnostic competence]. *Empirische Pädagogik*, 26, 292-311.

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

Lewis, Michael (ed.) (1997). *Implementing the Lexical Approach.* Hove: Language Teaching Publications.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, Vol. 23, No. 2, 5-12.

Nation, I., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle.

Rezaei, A., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.

Royal-Dawson, L., & Baird, J. (2009). Is Teaching Experience Necessary for Reliable Scoring of Extended English Questions? *Educational Measurement 28*/2, 2-8. doi: https://doi.org/10.1111/j.1745-3992.2009.00142.x

Schrader, F. (2013). Diagnostische Kompetenz von Lehrpersonen [Teachers' diagnostic competence]. *Beiträge zur Lehrerinnen und Lehrerbildung, 31*/2, 154-165.

Schütze, B., Souvignier, E. & Hasselhorn, M. (2018). Stichwort–Formative assessment. *Zeitschrift für Erziehungswissenschaft*, 21 (4), 679-715. https://doi.org/10.1007/s11618-018-0838-7

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76/1, 127-133.

Südkamp, A., & Praetorius, A. K. (Eds.). (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. [Teachers' diagnostic competences: theoretical and methodical advances]. Waxmann Verlag.

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. http://dx.doi.org/10.1037/a0027627

Visible Learning Plus (2018). *Visible learning plus. 250+ influences on student achievement*. Online: https://us.corwin.com/sites/default/files/250_influences_10.1.2018.pdf

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

Vögelin, C., Jansen, T., Keller, S., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: an analysis of teacher comments. *The Language Learning Journal.* doi:10.1080/09571736.2018.1522662

Weigle, S.C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.
Wolfe, E., & Jiao, H. (2016). Features of difficult to score essays. *Assessing Writing, 27*, 1-10.

Appendix A: Analytic rating scales used in text assessment

Frame of essay: Introduction and conclusion

4 Effective introduction with 'hook' and 'thesis statement'; effective conclusion summarizing main arguments

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

2 Introduction and/or conclusion identifiable but only partly effective

1 Both introduction and conclusion not clearly identifiable or mostly ineffective

Body of essay: Internal organisation of paragraphs

4 Paragraphs are well-organized and coherent throughout

3 Paragraphs are mostly well-organized and coherent

2 Paragraphs are partly well-organized and coherent

1 Paragraphs are not well-organized and incoherent

Support of arguments

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

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

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

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

Spelling and punctuation ('mechanics')

4 Author uses mostly correct spelling and punctuation

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

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

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

Grammar

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

DOI: https://doi.org/10.23770/pending

3 Author uses some complex grammatical structures, grammar mostly correct

2 Author uses few complex grammatical structures, grammar partly correct

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

Vocabulary

4 Author uses sophisticated, varied vocabulary throughout

3 Author mostly uses sophisticated, varied vocabulary

2 Author partly uses sophisticated, varied vocabulary, sometimes repetitive

1 Author uses little sophisticated, varied vocabulary, often repetitive

Overall task completion

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

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

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

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

Appendix B: Evaluation of videos

The following items were used to evaluate participants' perception of videos.

1. On a scale from 1-4, how comprehensible is the video?

   1 = incomprehensible, 4 = comprehensible

2. Did you have problems understanding the content of the video?
3. How useful do you find this video to assess vocabulary in students' texts? (1 = not useful at all, 4 = very useful)
4. The video was spoken…
   a. …too fast
   b. …at an appropriate speed
   c. …too slowly
5. The video covers…
   a. …too many theoretical constructs and examples
   b. …the right amount of theoretical constructs and examples
   c. …too few theoretical constructs and examples
6. I would have liked more of …
7. I would have liked less of …

DOI: https://doi.org/10.23770/pending

Appendix C: Manipulation check

The following items were used in the manipulation check.

a. Which are two main constructs of vocabulary?
   i. Lexical word fields and lexical diversity
   ii. Lexical sophistication and lexical standards
   iii. Lexical diversity and lexical sophistication
   iv. Lexical word fields and lexical standards

b. To assess vocabulary, it is particularly important to note the difference between vocabulary and…
   i. spelling, grammar and frame of essay
   ii. spelling, support of arguments and overall task completion
   iii. rhetoric, organisation and style
   iv. handwriting, spelling and formatting

c. What does the category frame of essay usually describe?
   i. Title, topic sentence and first argument of an essay
   ii. Writing prompt, genre and intended length of an essay
   iii. General topical framework in which the essay is placed
   iv. Introduction, body paragraphs and conclusion of an essay

Stefan D. Keller is professor of teaching and learning of English at the Institute of Secondary Education, School of Education, University of Applied Sciences and Arts Northwestern Switzerland. He is also deputy director of the Institute for Educational Sciences, University of Basel / PH FHNW.

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.

Nils Machts 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 assessment and use of rubrics in a wide range of educational contexts.

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

DOI: https://doi.org/10.23770/pending