Organisational quality of ESL argumentative essays and its influence on pre-service teachers’ judgments

Cristina Vögelin1*, Thorben Jansen2, Stefan D. Keller1, Nils Machts2 and Jens Möller2

Abstract: Assessing student writing is a complex task and experimental studies have shown that textual determinants, such as spelling or vocabulary, affect teachers’ judgments of other analytic criteria in student essays. Among these elements, organisation has not been extensively explored. This experimental study examines how organisational quality influences teacher judgments of ESL student texts in order to obtain a comprehensive overview of important determinants in the assessment process. Pre-service English teachers (N = 53) in Switzerland and Germany assessed four argumentative essays of two different overall text qualities in which two organisational features had been manipulated in terms of cohesive devices and paragraphing. Findings show that texts with higher organisational quality were assessed more positively concerning all analytic scales than texts with low organisational quality, indicating halo effects. In addition, an interaction effect was found between overall text quality and organisational quality, suggesting that the effect of organisational quality was stronger when overall text quality was low.

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

Thorben Jansen is a postdoctoral researcher 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.

PUBLIC INTEREST STATEMENT
Many pre-service teachers report that they do not feel adequately prepared for the assessment of student writing after having completed teacher education. Assessing student writing is a complex task and research has shown that many factors influence teachers’ judgments, such as spelling mistakes. This means that teachers assigned lower grades to texts containing many spelling mistakes than to the same text with few spelling mistakes, even when assessing organisation, argumentation or content. While the influence of spelling mistakes has been well researched, the authors of this article analysed if a similar influence of organisation can be found on the assessment of student writing. Pre-service teachers assessed the same student texts in different versions, once with clear paragraph breaks and suitable linking words and once without any paragraph breaks and few linking words. The aim of this study is to know more about the influence of text features on teachers’ judgments in order to increase the quality of assessment in the writing classroom.
This study sheds new light on the interaction between individual text features in the assessment of complex learner texts. Practical implications regarding essay assessment in teacher education are discussed.

**Subjects:** Education; Educational Research; Bilingualism / ESL; Secondary Education; By Subject; Newly Qualified Teachers; Teacher Training; Applied Linguistics; Language Teaching & Learning

**Keywords:** organisation; cohesive devices; paragraphing; second language writing; writing assessment

1. Introduction

A growing body of research has investigated the influence of specific factors on teacher judgments of written student compositions, including students’ ethnicity (Kaiser et al., 2016), teachers’ knowledge and experience (e.g. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Ready & Wright, 2011) and textual features, such as spelling and vocabulary (Birkel & Birkel, 2002; Scannell & Marshall, 1966; Vögelin et al., 2019). This present study focuses on pre-service teachers since there is an increasing awareness of the importance of teacher education also in writing assessment studies (Dempsey et al., 2009). It aims at generating valuable insights for research on assessing student writing and its determinants while also providing practical information on how to develop pre-service teachers’ diagnostic skills in teacher education programmes. Teaching and assessing student writing is an essential component of teachers’ professional competences (Baumert & Kunter, 2006; Edelenbos & Kubanek-German, 2004; Schrader, 2013). Many pre-service teachers in Switzerland and Germany report that they do not feel sufficiently prepared to assess students’ classroom performances when they enter the profession (Raun & Meier, 2007) and that this deficiency is particularly strong where written compositions are concerned (Porsch, 2010). While there is an increased focus on teachers’ diagnostic competence in general (Südkamp & Praetorius, 2017), subject-specific components of the process of diagnosis need to be investigated in more detail (Alderson et al., 2014) in order to educate pre-service teachers to assess students’ English writing as fairly and objectively as possible.

Although previous analyses investigating the influence of specific linguistic features on teacher judgments have advanced the understanding of the process of writing assessment and its determinants, they have primarily focused on spelling (mistakes), grammar, and vocabulary (Rezaei & Lovorn, 2010; Scannell & Marshall, 1966; Vögelin et al., 2019). For instance, Rezaei and Lovorn (2010) showed that pre-service teachers were affected by 20 structural, mechanical, spelling and grammar errors in well-written essays and assigned lower scores to these texts, even in criteria relating solely to content. Similarly, Vögelin et al. (2019) showed that pre-service teachers assessed student texts with a high level of lexical sophistication and diversity significantly more positively not only with regard to vocabulary but also concerning grammar and frame of essay.

This study focuses on the genre of argumentative essays since it is an essential component of learning and teaching English in many European countries, especially at upper-secondary level (Keller, 2013; Zhu, 2001). Command of argumentative essays requires—among other linguistic and rhetorical skills—skills in structuring arguments in order to persuade the readership with clearly organised and coherent paragraphs. For that reason, organisation plays an important role when writing argumentative essays (Hyland, 2008). Because of its importance, organisation is included in most analytic rating scales as a separate scale, such as the ESL Composition Profile (Jacobs et al., 1981), the Special Test of English Proficiency (STEP; Lumley, 2002), or the 6 + 1 Trait Model (Culham, 2003). Despite its importance in argumentative texts, empirical studies investigating argumentative essays with regard to organisation are few and far between.

In order to address this gap in the research, the present experimental study investigates the influence of organisational quality in argumentative essays on teacher judgments by minimizing
the interference of other confounding variables. Its outcomes should provide a more detailed insight into the intricate process of assessing ESL compositions and raise awareness of its complexity for (pre-service) teachers.

2. Theoretical background
This review of the relevant literature first discusses core theoretical concepts connected to organisation in writing such as coherence, cohesion, and metadiscourse. Second, it discusses the assessment of organisation in student writing and explores possible ways of measuring and manipulating organisation in student texts using specific features, such as paragraphing and cohesive devices. Third, the review examines the role of organisation in student writing in relation to the substantial body of literature exploring other text features that influence teacher judgments.

2.1. Cohesion, coherence and metadiscourse
From an educational perspective, an argumentative essay is typically organised in three main “stages” and associated “moves” (Hyland, 1990). The main stages consist of an introduction with a thesis statement, paragraphs with arguments, and a conclusion summarizing the main arguments. Following this structure, organisation is foregrounded in most ESL writing course books as a key aspect of writing essays (Hyland, 2008). For instance, the widely used course book Writers at Work: The Essay by Zemach and Stafford-Yilmaz (2008) provides different tasks and instructions for students to teach them how to write introductions, organise information in paragraphs, and compose an effective conclusion which will persuade readers. The basic structure of argumentative essays is thus well defined, but the conceptualisation of each component varies in the numerous academic contexts and research projects in which the genre is relevant (Kruse, 2013). While some researchers might regard the presence of thesis statements and topic sentences as central for evaluating organisation, others place their emphasis more on physical appearance such as paragraphing and the presence of cohesive devices when assessing organisation (Erdosy, 2004; Ruegg & Sugiyama, 2013). Organisation is a feature of writing quality with a particularly wide array of definitions, making it one of the most difficult aspects of an essay to assess (Erdosy, 2004; Freedman, 1979), particularly because it shares considerable overlap with other text features (Ruegg & Sugiyama, 2013). Three concepts which are closely related to organisation in student writing—coherence, cohesion and metadiscourse—are relevant in this context.

For a long time, the terms coherence and cohesion were used interchangeably (Yang & Sun, 2012) and previous literature has often employed ambiguous definitions (Knoch, 2007; Lee, 2002; Ruegg & Sugiyama, 2013). As studies in discourse analysis emerged, a shift in focus on ties between sentences occurred (Lee, 2002). Halliday and Hason (2013) define cohesion as the presence of textual ties between sentences in order to build connections between ideas and to aid the reader in following the text (see also Crossley & McNamara, 2009; Lee, 2002; Yang & Sun, 2012). While textual cohesion seems to be superficial (Yang & Sun, 2012), it is an essential aspect of successful and coherent language comprehension (Crossley & McNamara, 2009). Thus, a text with a more proficient and frequent use of cohesive cues is usually deemed more coherent and understandable (McNamara & Kintsch, 1996; Yang & Sun, 2012).

By contrast, the concept of coherence describes the semantic relations between ideas in a text (McNamara & Kintsch, 1996) and refers to the mental representations that readers derive from a text (Crossley et al., 2016; McNamara et al., 2010; Witte & Faigley, 1981; Yang & Sun, 2012). As it is not directly represented on the linguistic level, the concept is often regarded as abstract and fuzzy (Connor, 1990; Lee, 2002). Even though coherence and cohesion interact to a considerable extent, a cohesive text is not necessarily coherent (Ruegg & Sugiyama, 2013; Witte & Faigley, 1981). A text might include cohesive devices that link its sentences but lacks coherence or a clear purpose.

Related to cohesion and coherence, the term metadiscourse is a relatively new approach widely employed in discourse analysis and language education (Hyland, 2005). Coined by Harris in 1959,
several scholars, such as Vande Kopple (1985), Crismore et al. (1993), and Hyland (2005), have further developed this concept and presented taxonomies to classify an array of metadiscoursal features. Metadiscourse refers to “the linguistic material in texts, whether spoken or written, that does not add anything to the propositional content but that is intended to help the listener or reader organize, interpret, and evaluate the information given” (Crismore et al., 1993, p. 40). Typically employed as an umbrella term, metadiscourse encompasses an open-ended set of cohesive and interpersonal features which signal the writer’s intent and intends to guide a receiver through a text (Crismore et al., 1993; Hyland, 2005; Hyland & Tse, 2004). Typical examples are transitions, such as in addition, thus, or finally. These words do not contribute to topic development but contribute to achieving the communicative purpose of supporting the reader in understanding a text (Crismore et al., 1993; Hyland, 2005; Vande Kopple, 1997).

Further, metadiscourse is a characteristic of good ESL and native speaker student compositions (Cheng & Steffensen, 1996) and an important component of argumentative writing (Hyland, 1998). It is a crucial element of learning a second or foreign language since the degree of organising a text and guiding readers through it differs between languages (Hyland, 2005). Although the number of studies on metadiscourse has increased, its relation to cohesion and coherence remains uncertain. Jalilifar and Alipour (2007) state that metadiscourse influences cohesion, which is grammar-bound in nature, and does not account for coherence, which is meaning-bound. Contrary, Lee (2002) argues that metadiscourse markers account for a coherent text. This study focuses on textual devices which do not add to the propositional content of a text but help readers organizing and interpreting information, thereby contributing to a coherent text (Cheng & Steffensen, 1996; Crismore et al., 1993; Lee, 2002).

There are numerous denominations of cohesive devices, such as metadiscoursal features, metadiscourse markers, connectives, and cohesive ties (Crossley et al., 2016; Hyland, 2005; McNamara et al., 2014; Witte & Faigley, 1981). According to Halliday and Hasan (2013), conjunctive elements are not cohesive per se but presuppose the presence of other elements in the discourse. They distinguish between cohesive ties achieving between-sentence cohesion and cohesive ties achieving within-sentence cohesion (Halliday & Hasan, 2013). Whereas the latter—intrasentence cohesion—is primarily determined by grammatical structures of sentences, intersentence cohesion describes the use of cohesive ties creating cohesion across sentence boundaries (Bridge & Winograd, 1982). According to Halliday and Hasan (2013), cohesive devices across sentence boundaries are more prominent since they are the sole source of texture in contrast to cohesive devices within a sentence which are also structural relations (Halliday & Hasan, 2013). The effect is thus more striking and the meaning more obvious. Furthermore, these cues are local in nature, meaning that the cohesive devices relate to sentence-level cohesion, in contrast to global cohesion, and include explicit connections between sentences and paragraphs (Crossley et al., 2016; Halliday & Hasan, 2013).

This study defines cohesive devices as elements that create links between sentences which help readers connect and organise materials without contributing to propositional content (Hyland, 2005; Jalilifar & Alipour, 2007; Vande Kopple, 1997). This is in line with Ruegg and Sugiyama (2013) definition of cohesive devices and their adaptation of Hyland and Tse (2004) model of metadiscourse. Previous research investigating the relationship between cohesive devices and writing quality of ESL/EFL student texts showed somewhat conflicting results. Several studies found that essays rated higher contain more cohesive ties than essays rated lower (Chiang, 2003; Tejada et al., 2015; Witte & Faigley, 1981; Yang & Sun, 2012). By contrast, previous research with expert raters shows divergent results with regard to first language (L1) as well as second language (L2) writing (Crossley et al., 2015, 2016; Crossley & McNamara, 2012; Guo et al., 2013). According to these studies, local cohesion indices are either negatively or not at all correlated with expert judgments of text coherence and text quality.
2.2. Assessing organisation in student writing

Previous research focused predominantly on different text features that influence raters and teachers when assessing organisation (Freedman, 1979; Ruegg & Sugiyama, 2013). Freedman (1979) examined the influence of text features on teachers’ holistic assessment of student texts by rewriting four essays to be weaker or stronger in the four categories: content, organisation, sentence structure, and mechanics. Each of the four categories consisted of several features that were rewritten. For example, the manipulation of organisation involved differing paragraphing (appropriate paragraphing vs. three misparagraphings per 250-word page), ordering (logical ordering, keeping main ideas together, respecting “given-new” strategies vs. violating logical order, separating main ideas, violating “given-new” strategies) and transitioning (appropriate transitions vs. deleting transitions) (Freedman, 1979). This resulted in 96 essays (12 text variations on eight different topics), which were rated by 12 teachers using a 4-point holistic scale. Analyses of variance revealed that content and organisation affected the holistic ratings the most, in contrast to sentence structure and mechanics (Freedman, 1979). However, in the experimental variation of her study, the exact influence of one feature was difficult to determine since several aspects of one category were rewritten at once. This should be examined further.

In a recent study, Ruegg and Sugiyama (2013) investigated features that raters are sensitive to when assessing organisation in students’ argumentative essays. Their findings revealed that raters seem to attach more importance to the most visible aspects of organisation, such as the number of paragraphs and the number of cohesive devices, than to the deeper level aspects such as coherent flow of ideas. An essay with several clearly distinguished paragraphs appeared to raters as more organized than a text consisting of only one paragraph (Ruegg & Sugiyama, 2013). Similarly, Crossley et al. (2014) conducted an empirical evaluation of linguistic features measured by simple natural language processing (NLP) tools. They found that the number of paragraphs of 11th grade students’ essays correlated positively with human-rated essay scores ($r = .388$, $p < .001$). However, these findings have not yet been verified in an empirical study which investigates how organisational quality of student essays affect teacher judgments of other criteria by manipulating organisational features. Based on previous studies suggesting that raters base their judgment of organisation on par-amount and distinctive criteria and assigning them to separate scales. Previous studies have shown that the use of analytic scales is particularly relevant for pre-service teachers who are not used to assessing

2.3. Text features influencing teacher judgments

Reviewing various textual features and their influence on teachers’ assessment of student essays, previous literature reported the influence of spelling, grammar and vocabulary on teachers’ analytic assessment of other criteria (Birkel & Birkel, 2002; Rezaei & Lavor, 2010; Scannell & Marshall, 1966; Vögelin et al., 2019). While Birkel and Birkel (2002) conducted their study with experienced elementary school teachers, many subsequent studies examined pre-service and in-service teachers’ judgments. For instance, pre-service teachers assigned lower grades to papers with punctuation, spelling, and grammar errors even though they were instructed to grade content alone (Scannell & Marshall, 1966). Furthermore, teachers are strongly influenced by the physical appearance of a text, for example, a student’s handwriting (Charney, 1984; Marshall & Powers, 1969). This refers to distortion effects in an analytic assessment situation, distinguishing between construct-relevant and construct-irrelevant factors (Messick, 1994). Teachers should be able to judge the construct-relevant factors of a student’s written performance (e.g., vocabulary, organisation, spelling, etc.) separately from construct-irrelevant factors such as the students’ ethnicity, or gender (Kaiser et al., 2016). Failing to distinguish between construct-relevant and construct-irrelevant factors indicates a halo effect (Thurstone, 1920). A halo effect describes the positive or negative influence of one feature on other independent criteria in the rating process (Bechger et al., 2010; Thorndike, 1920). In ESL writing assessment, teachers typically decide on a set of relevant analytic scales for a particular writing task or context. Behind this selected set of criteria, there is a much larger set of criteria applicable to describe writing quality which interacts and correlate with each other (Sadler, 1989). For instance, poor vocabulary skills might affect the quality of grammar in a student text and it might be difficult to differentiate between these individual dimensions of a text. In an analytic approach, teachers try to disambiguate these dimensions by selecting a limited number of relevant and distinctive criteria and assigning them to separate scales. Previous studies have shown that the use of analytic scales is particularly relevant for pre-service teachers who are not used to assessing
authentic, multifaceted student performances with uneven learning profiles (Weigle, 2002). In particular, second language writers who struggle with formal aspects of writing, such as spelling, benefit from analytic assessment, as different dimensions of writing will be judged separately. When regarding criteria separately, teachers are concerned with individual properties—or qualities—of a text (Sadler, 1989). This approach contrasts with a holistic approach in which teachers react to a piece of writing as a whole and do not assume operational independence among criteria, thus judging its quality in a broader sense. In a holistic approach, teachers make judgments based on the most salient criteria in a text and it might be difficult for students to make systematic progress if they do not know what configuration of criteria teachers are referring to in their appraisal of the text (Sadler, 1989). This is different in the analytic approach where students see which criteria are relevant for task completion, how a teacher rates each individual one, and consequently, in which areas they need to improve. Thus, it becomes an important aspect of teacher professionalism to be able to evaluate selected criteria relevant for a genre separately in an analytic assessment (Parr & Timperley, 2010), even though there might be some overlap between these aspects from a linguistic perspective.

In general, linguistic resources—in this case cohesive devices—have the function of displaying certain moves in a text (Swales, 1990). In the case of argumentative essays, cohesive devices typically display the order of arguments, which helps the reader to quickly grasp the structure of a text and thus follow it more easily. Another text pattern would be paragraphing for each argument. These are requirements of the genre argumentative essays and need to be learned before writing an essay. In this study, learners studied the structure of an argumentative essay and identified the different moves in the preceding teaching unit (cf. “Student texts”).

3. Purpose
Assessing writing in authentic classroom contexts entails an interplay of different influencing factors as well as a larger variety of text qualities. Features such as grammar, vocabulary or argumentation are distributed in varying configurations in different essays. This study, however, does not aim at representing the full diversity of students’ written compositions. The aim of this experimental study is to explore how the organisational quality of student texts influences pre-service teachers’ analytic assessment. Based on Ruegg and Sugiyama (2013) findings, we varied organisational quality in the student essays by manipulating two key features of organisation in ESL argumentative essays: the number of paragraphs and the number of cohesive devices. By manipulating organisational quality in ESL argumentative essays, we investigated whether the experimental variation of organisation affected teacher judgments of other criteria. Consequently, the current study addresses the following research question:

- Does the number of paragraphs and selected cohesive devices in ESL argumentative essays affect pre-service teachers’ judgments of distinct and independent text features (indicating halo effects)?

Based on the studies summarized above, and our own work on the influence of spelling and lexical quality on other features of text assessment, we hypothesise that organisation has a similar effect and influences teacher perception of text features which should—from a pedagogical perspective—be viewed as separate.

4. Method
This study examines organisation from an educational perspective within the context of upper-secondary ESL writing. In order to obtain detailed insights into the influence of specific textual determinants in the assessment process, this study employs a similar design and methodology as an earlier study which investigated the influence of lexical diversity and sophistication on teacher judgments (Vögelin et al., 2019). This new, independent study involved the variation of organisational quality and was conducted with a different sample of participants.
4.1. Student texts

First, a teaching unit on argumentative writing was developed and implemented at an upper-secondary grammar school (“Gymnasium”) in Switzerland. This type of schooling (International Standard Classification of Education [ISCED] level 3) is quite selective in Switzerland: Only about 21% of students completed their upper-secondary education in this type of school in 2015 (National Statistical Office of Switzerland, 2016). Over 4 weeks, their English teacher taught them techniques of argumentative writing with materials from the coursebook Writers at Work—The Essay (Zemach & Stafford-Yilmaz, 2008). The teaching unit encompassed watching a video on “how technology shapes our life” with subsequent discussion, reading various articles on the influence of technology on people’s lives, analysing the structure of argumentative essays, finding examples for one’s arguments and independent writing of parts of argumentative essays. Thus, students learned how to structure and organize different elements of an argumentative essay, how to sustain their arguments by providing different types of support and by using the appropriate type of language required by the genre. At the end of 4 weeks, students were asked to write an essay in 90 minutes answering the following writing prompt: “Do you agree or disagree with the following statement? As humans are becoming more dependent on technology, they are gradually losing their independence.” The students were in 11th grade (2 years before their final examination), had been studying English for 5 years, and were upper-intermediate learners.

From this dataset of 15 ESL argumentative essays, two texts with a high overall text quality and two texts with a low overall text quality were selected using the NAEP rating scale (National Assessment Governing Board, 2010). While the weaker student compositions exhibited insufficient levels of work with regard to the learning goals of the unit, the stronger compositions showed levels of work that were considered to have surpassed the learning goals. After preselecting two strong and two weak texts, two groups of independent expert raters (N = 9 and N = 7), consisting of experienced English lecturers and teachers that were part of the research team, rated one weak and one strong text, respectively, upon their overall text quality using the NAEP rating scale. The intra-class correlation coefficients (ICC) were satisfying with ICC (2,9) = .88 and ICC (2,7) = .75.

As outlined in the literature review, previous studies have reported that raters base their assessment of organisation in student writing on paragraphing and the number of cohesive devices. Since both features are central skills for upper-secondary ESL students to write argumentative texts, this study experimentally varied these features in the four student texts to investigate their role in the assessment process. We are aware that, from a linguistic point of view, our selected features do not cover the entire construct of organisation. We chose this operationalisation because first, paragraphing and cohesive devices are viewed as central for writing well-structured argumentative essays at upper-secondary level in the pedagogical literature and feature heavily in relevant course-books. Second, empirical studies have shown these two features to be particularly salient in determining raters’ view of organizational quality in student texts (Ruegg & Sugiyama, 2013).

The study defined cohesive devices following the model of metadiscourse by Hyland and Tse (2004) and the study by Ruegg and Sugiyama (2013). In addition, we analysed cohesive devices found in our dataset of authentic student essays and complemented the list with further devices of our own. Table 1 displays the selection of cohesive devices employed in this study. In order to ensure that the manipulated number of cohesive devices was within the range of authentic student essays at that level, we examined the number of features per 100 words in the ESL compositions of the data set. We found a minimum of 0.75 tokens and a maximum of 2.80 tokens of cohesive devices per 100 words. Based on the mean length of M = 459.75 words of the four essays, we included three cohesive features in essays with a low organisational quality and 12 features in essays with a high organisational quality. Depending on the cohesive devices present in the original essay, different cohesive devices were inserted or substituted in order to increase coherence. To standardize the number and type of cohesive devices between texts, some cohesive devices already present in a student text had to be deleted or altered at times. All cohesive devices
were inserted at the beginning of a sentence and carried little or no propositional meaning, as in the following example: “In addition, you cannot lose time if you use your phone” (own emphasis). In texts with low organisational quality, connectives consisted only of the tokens so, because (of), and and.

There were 12 tokens (11 types) of cohesive devices in each text with high organisational quality: six transitions, four frame markers, one code gloss, and one attitude marker. The level of lexical sophistication and diversity was measured between the text variations using Coh-Metrix (Graesser et al., 2004) and the Tool for the Automatic Analysis ofLexical Sophistication (TAALES) (Kyle & Crossley, 2015). Using these objective measures, we made certain that the insertion of sophisticated cohesive devices in texts with a mean length of $M = 459.75$ words did not lead to a significant difference in the overall quality of vocabulary.

Regarding paragraphing, we deleted all paragraph breaks in texts with the low organisational quality so that the essay appeared as one large paragraph. In texts with high organisational quality, essays were structured into five or more paragraphs according to the existing arguments in the text (i.e., each paragraph contained one idea or argument).

We thus manipulated organisational quality by systematically varying both paragraphing and cohesive devices in four student texts. This created a corpus of eight texts in which each text—regardless of overall text quality—existed in two versions: one with low organisational quality and one with high organisational quality. These texts then formed the basis of our experimental study.

### 4.2. Rating scales

This study employed both analytic and holistic rating scales (Crusan, 2010; Knoch, 2009; Rakedzon & Baram-Tsabari, 2017; Weigle, 2002) to investigate whether organisational quality influences teacher judgments on both types of scales. To measure the influence of organisation on other features of text quality, and to address the genre-specific characteristics of argumentative essays, we developed specific rating scales aligned with the relevant syllabus of the 11th grade at the upper-secondary grammar school in Switzerland and Germany. Our scales were based on existing analytic scales where appropriate, such as the 6 + 1 trait model (Culham, 2003) and the Test in English for Educational Purposes (TEEP; Weir, 1988). This resulted in the following six analytic scales: organisation, support of arguments, spelling and punctuation, grammar, vocabulary, and overall task completion (see Appendix). Special attention was payed to the organisational scale, since previous studies reported raters’ difficulties in understanding the descriptors of this concept (Ruegg & Sugiyama, 2013) or their tendency to follow their own internal manifestations of organisation (Erdosy, 2004). Thus, the descriptors of the

| **Table 1. Categorised cohesive devices** |
|-----------------------------------------|
| **Category**   | **Examples** |
| Transitions   | additionally, in addition, also, and, because, besides, but, due to that fact, further, furthermore, however, moreover, nevertheless, nonetheless, on the one hand, on the other hand, so, therefore, thus, well, yet |
| Frame markers | all in all, finally, to finish, first, first of all, in conclusion, last but not least, last, next, overall, second, to sum up, to summarize, to conclude |
| Time markers  | nowadays, then, these days, today, now |
| Code glosses  | for example |
| Boosters      | of course, clearly, in fact, surely |
| Attitude markers | (in) my opinion, personally, for me |

Note: Adapted from Ruegg and Sugiyama (2013) and Hyland and Tse (2004).
organisational scale are based on the **KEPT Essay Rating scales** (Ruegg & Sugiyama, 2013), the **ESL Composition Profile** (Jacobs et al., 1981) and the **STEP writing performance criteria and descriptors** (Lumley, 2002). In a pilot phase, a group of experienced lecturers created an a priori rating scale together with the research team, in which they adapted and evaluated the assessment criteria in two rounds of feedback in order to eliminate imprecise wording and ambiguities. As previous studies have shown no significant differences between 4-point and 6-point rating scales with regard to reliability and sensitivity (McColloy & Remsted, 1965; Nimehchisalem, 2010), we chose rating scales with only four levels in order to make them practical and easy to use.

We also included a holistic scale in this study as a manipulation check. A successful manipulation would include a lower holistic assessment for texts with lower organisational quality, and vice versa, since organisation is a component of the holistic scale employed in this study. We used the well-validated National Assessment of Educational Progress (NAEP) rating scale, which is divided into six levels (National Assessment Governing Board, 2010). This scale was previously used in large-scale empirical writing assessment studies in Germany because it contains age and grade appropriate criteria fitting well with the national curriculum (Knopp et al., 2012). In the Swiss and German educational context, no standardized holistic rating scale for argumentative essays at upper-secondary level exists. Thus, we chose the NAEP scale as it is widely used in international research and proves a good fit with local curricula.

Existing theoretical work on the nature of analytic assessment has shown that the criteria used in appraising the quality of performance interlock to some degree (Bacha, 2001; Sadler, 1989). Analytic criteria may sometimes be fuzzy or be regarded as a continuous gradation from one category to another, as in assessing “originality” (Sadler, 1989). In designing our own scales, we, therefore, paid special attention to the distinctiveness of our genre-specific analytic rating scales by differentiating between the six scales and defining the levels within each scale by using detailed descriptors. Our aim was to test whether participants would be able to discern these criteria as a profile of distinct strengths and weaknesses, and to gauge the size of a possible influence of our manipulated criterion.

### 4.3. Participants

Participants were students attending seminars on English language methodology or English language skills at Swiss and German universities (N = 53). 60.4% of the participants were enrolled in a teacher’s education programme for upper-secondary school and 28.3% in a programme for lower-secondary school. The remaining 11.3% were enrolled in an English master’s degree with the intention of later becoming upper-secondary teachers. The participants’ age ranged from 21 to 48 years with a mean of 27.98 years (SD = 5.78). The majority of participants were female (69.8%). Furthermore, 54.7% of the participants reported that their English proficiency corresponded to level C2 according to the *Common European Framework* (Council of Europe, 2001). The remaining participants reported that their English was equivalent to a C1 (35.8%) or that English was their first language (7.5%). The study thus dealt with participants who, in their own estimate, were highly competent users of the English language.

### 4.4. Procedure

This study used the computer-based assessment tool **Student Inventory ASSET** (SIA) which has been adapted from the Student Inventory (SI) (Vögelin et al., 2019). First, participants received background information on the genre of the texts (argumentative essays), the school context, and the preceding teaching unit which had been implemented prior to the writing prompt. After reading this background information, participants were asked to familiarize themselves with the descriptors of the holistic and analytic rating scales employed in the study. Each participant then evaluated four texts in randomized order—two weak texts with differing levels of organisational quality (high/low) and two strong texts with differing levels of organisational quality (high/low). Participants assessed each text on the holistic scale first and then on the analytic rating scales. In
order to control for the influence of handwriting (Marshall & Powers, 1969), student essays were typed and displayed in the same font. Last, participants were asked background questions concerning their studies and English language proficiency. After data gathering was complete, participants saw their scores in relation to expert ratings of the four texts. In addition to the expert ratings on both types of scales, experts’ descriptive comments on the student texts were presented in order to justify their ratings. In that way, participants could see where they had assessed more strictly or leniently and in what ways their reasons for the assessment aligned with that of the experts. This feedback was not part of the empirical study but was added at the end as a benefit to participants who had taken part in the study.

4.5. Analyses

We used an experimental $2 \times 2$ design with two independent variables: overall text quality and organisational quality. To test the hypothesis, we conducted repeated-measures multivariate analyses of variance, which assume normal distribution of the data, independence, and sphericity. Concerning the first assumption, the analysis of variance is robust against the violation of normal distribution since our experimental design was balanced with more than 20 participants per condition (Schmider et al., 2010). Further, the second and third assumptions were given by our design. Table 2 displays skewness and kurtosis values, which were in acceptable ranges for all dependent variables.

There were no missing values as participants could only proceed in SIA once they had assessed all four texts. We calculated intra-class coefficients (ICC) for the inter-rater agreement on individual texts. Inter-rater agreement ranged from ICC (2,1) = .34 on the scale task completion and ICC (2,1) = .74 on the scale organisation. The relatively low ICC coefficients on scales other than organisation indicate the substantial variation of ratings of each of the eight texts in comparison to the variation of ratings across the texts. Yet, the eight texts constitute twice the same four texts with a slight variation of organisation and only two variations of overall text quality across the four texts. Therefore, we can safely assume that the low ICCs are a result of the low overall variance of text quality and should expect to receive high ICCs only on the ratings of organisation.

5. Results

5.1. Overall text quality

Table 3 shows descriptive statistics for the variable text quality with regard to the holistic and the six analytic rating scales. Correlations between the holistic and analytic scores ranged between $r = .61 \ (p < .001)$ for spelling and $r = .74 \ (p < .001)$ for support of arguments (see Table 4). The

| Table 2. Descriptive statistics for teachers’ holistic and analytic ratings over all texts |
|---|
| Min | Max | Mean (SD) | Skewness (SE) | Kurtosis (SE) | ICC |
|---|
| Holistic scale | 1 | 6 | 3.83 (1.28) | −.06 (0.17) | −.65 (0.33) | 0.51 |
| Analytic ratings | 1 | 4 | 2.61 (1.15) | −.09 (0.17) | −1.43 (0.33) | 0.74 |
| Organisation | 2.70 (0.81) | −.17 (0.17) | −.43 (0.33) | 0.37 |
| Support of arguments | 1 | 4 | 2.77 (0.91) | −.23 (0.17) | −.77 (0.33) | 0.57 |
| Spelling | 2.66 (0.91) | −.14 (0.17) | −.78 (0.33) | 0.54 |
| Grammar | 2.67 (0.96) | −.20 (0.17) | −.93 (0.33) | 0.51 |
| Vocabulary | 2.68 (0.84) | −.02 (0.17) | −.68 (0.33) | 0.34 |

Vögelin et al., Cogent Education (2020), 7: 1760188
https://doi.org/10.1080/2331186X.2020.1760188
multivariate test indicated significant effects for text quality \( (F(7, 46) = 44.41, p < .001, \eta^2 = .87) \) and thus, we conducted univariate post-hoc tests.

The first set of univariate tests showed that texts with a high overall text quality were judged more positively than texts with a low overall text quality with regard to the holistic score \( (F(1, 52) = 128.91, p < .001, \eta^2 = .71) \). Significant effects were further found for all analytic criteria: organisation \( (F(1, 52) = 48.51, p < .001, \eta^2 = .48) \), support of arguments \( (F(1, 52) = 73.51, p < .001, \eta^2 = .59) \), spelling \( (F(1, 52) = 187.36, p < .001, \eta^2 = .78) \), grammar \( (F(1, 52) = 234.40, p < .001, \eta^2 = .82) \), vocabulary \( (F(1, 52) = 176.77, p < .001, \eta^2 = .77) \), and task completion \( (F(1, 52) = 48.12, p < .001, \eta^2 = .48) \). Hence, texts with a high overall text quality were judged more positively on all analytic scales than texts with a low overall text quality.

### 5.2. Organisational quality

Concerning the variable organisational quality, the multivariate test indicated significant effects \( (F(7, 46) = 46.99, p < .001, \eta^2 = .88) \); thus, we conducted univariate post-hoc tests (see Table 5).

---

**Table 3. Teachers’ holistic and analytic ratings of texts with low and high text quality.**

|                  | Low text quality | High text quality | Mean difference | d  |
|------------------|------------------|-------------------|-----------------|----|
|                  | M    | SD  | M    | SD  |                |     |
| Holistic scale   | 3.13 | 0.73| 4.53 | 0.77| 1.40*          | 1.87|
| Analytic ratings |      |     |      |     |                |     |
| Organisation     | 2.36 | 0.38| 2.87 | 0.49| 0.51*          | 1.16|
| Support of arguments | 2.33 | 0.54| 3.07 | 0.45| 0.74*          | 1.49|
| Spelling         | 2.15 | 0.52| 3.40 | 0.46| 1.25*          | 2.55|
| Grammar          | 2.07 | 0.50| 3.25 | 0.46| 1.18*          | 2.46|
| Vocabulary       | 2.18 | 0.58| 3.29 | 0.46| 1.11*          | 2.12|
| Task completion  | 2.31 | 0.52| 3.06 | 0.55| 0.75*          | 1.40|

Note: * p < .001

**Table 4. Correlations between rating scales.**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1) Holistic rating                  |   |   |   |   |   |
| 2) Organisation                    | .62|   |   |   |   |
| 3) Support of arguments             | .74| .44|   |   |   |
| 4) Spelling                         | .61| .38|.52|   |   |
| 5) Grammar                          | .66| .47|.57|.69|   |
| 6) Vocabulary                       | .68| .53|.62|.66|.75|
| 7) Task completion                  | .70| .42|.68|.41|.46|.51|

Note: All correlations are significant \( (p < .01) \).
The second set of univariate tests showed that texts with a high organisational quality were judged more positively than texts with a low organisational quality with regard to the holistic score ($F(1, 52) = 52.00, p < .001, \eta^2 = .50$), indicating a successful manipulation check.

Turning to the analytic scale, significant effects of organisational quality were found for all analytic criteria. The strongest effect was—as expected—on organisation ($F(1, 52) = 295.37, p < .001, \eta^2 = .85$), followed by vocabulary ($F(1, 52) = 35.09, p < .001, \eta^2 = .40$), task completion ($F(1, 52) = 29.96, p < .001, \eta^2 = .37$), support of arguments ($F(1, 52) = 28.53, p < .001, \eta^2 = .35$), grammar ($F(1, 52) = 16.74, p < .001, \eta^2 = .24$), and spelling ($F(1, 52) = 13.16, p = .01, \eta^2 = .20$). Texts with high organisational quality were judged more positively on all analytic scales than texts with low organisational quality.

As we will discuss below, these results are indicative of halo effects for certain text features.

### 5.3. Interaction between text quality and organisational quality

A multivariate effect was detected for first-level interactions between text quality and organisational quality ($F(7, 46) = 2.36, p < .05, \eta^2 = .26$). The third set of univariate tests for the interaction effect indicated that the effect of organisational quality was stronger when overall text quality was low. The effects of organisation on the judgment of organisation ($F(1, 52) = 5.57, p < .05, \eta^2 = .10$) and support of arguments ($F(1, 52) = 6.34, p < .05, \eta^2 = .11$) was higher in texts of low overall quality than in texts of high overall quality (see Figures 1 and 2).

### 6. Discussion

The aim of this experimental study was to explore how organisational quality of ESL argumentative essays influences pre-service teachers’ judgments of other independent textual features in an analytic assessment process. By focusing on organisation—operationalised by paragraphing and cohesive devices—as an influencing factor on teacher judgments, this study provides new insights into the assessment process of ESL texts which—we hope—are beneficial both for the research and practice of writing assessment. We focused on pre-service teachers in this study because there is an increasing awareness of the importance of teacher education for developing diagnostic competence in general (Blömeke et al., 2015). Especially in writing education, it has been shown that teaching experience does not automatically transfer to high diagnostic competence or accurate assessment. For example, Jansen et al. (submitted) 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).

### Table 5. Teachers’ holistic and analytic ratings of texts with low and high organisational quality

|                      | Low organisational quality | High organisational quality | Mean difference | d  |
|----------------------|---------------------------|-----------------------------|-----------------|----|
| **Holistic scale**   | 3.33 0.73                 | 4.33 0.84                   | 1.00*           | 1.27 |
| **Analytic ratings** |                           |                             |                 |    |
| Organisation         | 1.71 0.45                 | 3.52 0.58                   | 1.81*           | 3.49 |
| Support of arguments | 2.65 0.50                 | 2.94 0.52                   | 0.49*           | 0.96 |
| Spelling             | 2.61 0.52                 | 2.93 0.45                   | 0.32*           | 0.66 |
| Grammar              | 2.45 0.52                 | 2.87 0.55                   | 0.42*           | 0.79 |
| Vocabulary           | 2.42 0.56                 | 3.05 0.59                   | 0.63*           | 1.10 |
| Task completion      | 2.43 0.50                 | 2.93 0.49                   | 0.50*           | 1.01 |

Note: * $p < .001$. 

The second set of univariate tests showed that texts with a high organisational quality were judged more positively than texts with a low organisational quality with regard to the holistic score ($F(1, 52) = 52.00, p < .001, \eta^2 = .50$), indicating a successful manipulation check.
in comparison to benchmarks scores. As Dempsey et al. (2009) showed, learning to assess student texts objectively and fairly requires scaffolded practice in assessing multiple student papers, applying explicit assessment criteria under expert guidance and discussing assessments in peer groups. It is this type of specific training that pre-service teachers need to receive in their teacher education, and our study aimed to lay the groundwork for suitable materials and training tools in this field.

Findings suggest that pre-service teachers were able to differentiate between strong and weak ESL argumentative essays on holistic as well as analytic rating scales, replicating results from a previous study on ESL text assessment (Vögelin et al., 2019). Thus, participants judged texts with higher overall text quality more positively than texts with lower overall text quality with regard to all criteria. In order to explore the role of organisation on other analytic scales, this study further investigated whether experimentally varied organisational qualities of student texts influence teachers’ judgments of other analytic criteria. Results showed that the manipulation of organisational quality significantly affected participants’ assessment on a holistic scale, indicating a successful manipulation check. Furthermore, the variation of organisational quality also influenced assessment on all analytic scales employed in this study: support of arguments, spelling,
grammar, vocabulary, and overall task completion. On all these features, texts with a higher organisational quality were rated more positively than texts with a lower organisational quality.

While findings indicate halo effects for criteria which are sufficiently distinct from organisation, other criteria might be more difficult to distinguish from organisation in student texts. In our view, the influence of organisation on formal criteria, such as spelling and grammar, suggests halo effects. Even though the number of spelling mistakes was kept constant between text variations, spelling was assessed more positively in texts with higher organisational quality. The same can be said for grammar, which was defined as usage of a variety of complex grammatical structures and few grammatical mistakes. Again, organisation should also not affect the assessment of grammar since cohesive devices were inserted solely at the beginning of a sentence and did not alter the original grammatical structure of the student essay in any way.

We also argue for halo effects for the criterion overall task completion as it was operationalized in this study. According to our analytic rating scale, task completion was defined as whether the student text fully addresses the essay question and was thus concerned whether the written text is relevant to the essay question (see Appendix). As the variety of transition words and the paragraphing of a text only marginally overlap with the criterion of whether an essay fully addresses the essay questions, we would again argue that this constitutes a halo effect.

The criteria support of arguments and vocabulary are more complex cases since they share some overlap with the analytic scale organisation. First, the criterion support of arguments was defined in this study as to whether authors present a variety of different examples to support their argument and whether the relevance of these examples to the topic is fully explained. By inserting cohesive devices, the relevance of examples might be better explained to the readership. However, the influence of one criterion on the other is still remarkable as students’ abilities to present arguments and clarify their relevance to the topic are noticeably different from their use of paragraphing and cohesive devices which do not add to the propositional content of the essay.

Second, quality of vocabulary—defined as whether the author uses sophisticated, varied vocabulary—correlates operationally with organisation since the insertion or deletion of cohesive devices affects the lexical quality of the texts. Further, the two scales share some overlap with regard to the formulation “sophisticated transition words” and “sophisticated, varied vocabulary”. This should be avoided in future studies. Nevertheless, the automated tools of text analysis (Coh-Metrix and TAALES) employed in this study indicated that lexical diversity and sophistication stayed within comparable range, and the differences perceived by human raters in our experimental variation far exceeded the minor changes made in our manipulation.

Overall, the findings of this study suggest that the manipulation of organisational quality affected the assessment of other analytic scales to a greater and lesser degree. While the effect of organisation on the analytic scales spelling, grammar, and task completion indicates halo effects in pre-service teachers’ assessment, the influence on vocabulary and support of arguments is not that clear-cut due to overlap between the analytic criteria. In sum, these findings underscore how difficult it is for pre-service teachers to keep different categories in analytic rating scales apart and assess each correctly. Findings further support the notion that teachers partly base their assessment—consciously or subconsciously—on the most noticeable aspects of writing (Charney, 1984; Erdosy, 2004; Ruegg & Sugiyama, 2013), since the successful manipulation of organisational quality in this study focused on the most visible aspects of organisation (paragraphing) and on cohesive devices only. Overall, results reinforce those reported in previous studies, which is that analytic assessments are prone to halo effects which can lead to unfair assessments of student writing (Birkel & Birkel, 2002; Freedman, 1979; Rezaei & Lovorn, 2010; Scannell & Marshall, 1966; Vögelin et al., 2019, 2018). Our study suggests that organisation should play an important role in ESL teacher education as it influences pre-service teachers’ judgments of other criteria in significant ways.
7. Implications

7.1. Implications for practice
This study contributes one modest piece to the large puzzle of text assessment involving the multifaceted construct of organisation, which is highly relevant for teachers of ESL argumentative writing at upper-secondary level. Results imply a need for raising awareness among pre-service teachers of possible influences in the assessment process and preparing them to assess student texts as fairly and objectively as possible by using analytic rating scales. Pre-service teachers should be aware of the influence of cohesive devices and paragraphing on their assessment of other criteria, such as spelling, grammar and task completion. They should further be familiarized with key elements of organisation in argumentative essays in order to teach students how to present and signal their arguments in a successful argumentative essay. Our study illustrates that it is important to discuss and practice assessing individual text features separately on analytic scales in teacher education. This could be realized in an initial practice session in which pre-service teachers could assess multiple authentic student texts in SIA and receive immediate feedback. Pre-service teachers could discuss in peer groups their assessment experiences and the influence of specific text features. In a subsequent session on rating scales, participants could discuss the advantages and disadvantages of various rating scales and apply them under expert guidance. This is particularly relevant for Swiss and German pre-service teachers who can later choose their own rating scales for use in their classroom assessments in English writing.

Further research should also develop subject-specific analytic rating scales which are aligned with the local syllabi of the educational contexts in which they are used (Rakedzon & Baram-Tsabari, 2017; Rezaei & Lovorn, 2010). Subject-specific rating scales enable teachers to identify, examine and credit specific elements of a genre more easily in comparison to general rating scales.

7.2. Limitations and directions for future research
This study employed a controlled manipulation of two organisational features within an experimental setting and thus follows a narrow operationalisation of the broad construct of organisation in student writing. As it did not aim at covering the complete range of possible text variations relating to organization which occur in authentic student texts, a generalization to a real assessment situation would require additional studies.

While our operationalization of organisation was somewhat limited, the manipulation employed in this study resembled naturally occurring phenomena and reflected the nature as well as the frequency of organisational features within the range of student texts in the dataset upon which this research is based. However, future studies should investigate the influence of organisational quality on the basis of a larger corpus of student essays in order to cover the wide range of quality occurring within authentic written performances.

This study is further limited due to the combined manipulation of cohesive devices and paragraphing. This co-occurrence of manipulated text features does not allow any conclusions about which feature had more influence on teacher judgments. Following previous research stating that teachers are strongly influenced by superficial aspects of student writing (Charney, 1984; Scannell & Marshall, 1966), one might assume that the variation of paragraphs had a greater influence on teachers’ assessment, since it is more prominent at first glance. It would be interesting to analyse the precise influence of each text feature in separate studies. We included both features in this first study as they are essential components of organisation from a rater’s perspective (Ruegg & Sugiyama, 2013).

Further, this study defined organisational quality of a student text as surface level indicators. In contrast to previous studies that investigated the underlying constructs of organisation, such as
cohesion and writing proficiency (Crossley et al., 2014; Crossley & McNamara, 2012) or coherence (Lee, 2002; McNamara et al., 2010; Witte & Faigley, 1981), we intended to explore the influence of two organisational features on teacher judgments in an experimental setting. Manipulating a feature such as coherence would have implied complete re-writing of original texts, inevitably also changing aspects such as grammar or vocabulary which were distinct variables in our design. Still, our setting is not completely artificial as our dataset of argumentative essays exhibited several compositions by students who seemed to be struggling with appropriate paragraphing or effective use of cohesive devices.

Future studies could explore how halo effects can be minimized in order to make writing assessment more objective. They could also investigate teacher judgments in a more “natural” context. How do pre-service and in-service teachers assess organisation in classroom assessment activities? Additionally, it would be interesting to examine the potential of teacher training programmes? These studies could also compare first and second/foreign language writing assessment.

In conclusion, we believe that our study enriches the comprehension of pre-service teachers’ assessment of ESL essays and contributes useful data on an essential aspect of diagnosis in the English classroom. The ultimate goal of this research is to contribute to more precise and fairer assessments of student performances in the English classroom.

Funding
This work was supported by the Swiss National Science Foundation under Grant 165483; and the German research foundation (DFG) under Grant Ma648/25-1.

Author details
Cristina Vögelin
E-mail: cristina.voegelin@unibas.ch
ORCID ID: http://orcid.org/0000-0003-0115-5314

Thorben Jansen
E-mail: tjansen@ipl.uni-kiel.de
ORCID ID: http://orcid.org/0000-0003-1767-5859

Nils Machts
E-mail: nmachts@ipl.uni-kiel.de

Jens Möller
E-mail: jmoeller@ipl.uni-kiel.de
ORCID ID: http://orcid.org/0000-0002-5621-7570

Institute for Educational Sciences, University of Basel, Muttenz, Switzerland.

Institute of Psychology of Learning and Instruction, Kiel University, Kiel, Germany.

Citation information
Cite this article as: Organisational quality of ESL argumentative essays and its influence on pre-service teachers’ judgments, Cristina Vögelin, Thorben Jansen, Stefan D. Keller, Nils Machts & Jens Möller, Cogent Education (2020), 7: 1760188.

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. https://doi.org/10.1093/applin/amt046
Bach, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? System, 29(3), 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2
Bates, C., & Nettelbeck, T. (2001). Primary school teachers’ judgements of reading achievement. Educational Psychology, 21(2), 177–187. https://doi.org/10.1080/01443410020043878
Baumert, J., & Kunter, M. (2006). Keyword: Professional competencies of teachers. Zeitschrift Für Erziehungswissenschaften, 4(9), 469–520. https://doi.org/10.1007/s11618-006-0165-2
Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. Applied Psychological Measurement, 34(8), 607–619. https://doi.org/10.1177/014662161037897
Birkel, P., & Birkel, C. (2002). How concordant are teachers’ essay scorings? A replication of Rudolf Weiss’ studies. Psychologie in Erziehung Und Unterricht, 49 (3), 219–224.
Blomme, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. Zeitschrift Für Psychologie, 221(1), 3–13. https://doi.org/10.1027/2151-2604/a000194
Bridge, C. A., & Winograd, P. N. (1982). Readers’ awareness of cohesive relationships during cloze comprehension. Journal of Reading Behavior, XIV(3), 299–312. https://doi.org/10.1080/01443410020043878
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.
Cheng, X., & Steffensen, M. S. (1996). Metadiscourse: A technique for improving student writing. Research in the Teaching of English, 30(2), 149–181.
Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. System, 31(4), 471–484. https://doi.org/10.1016/j.system.2003.02.002
Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. Research in the Teaching of English, 24(1), 67–87.
Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.
Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing. Written Communication, 10(1), 39–71. https://doi.org/10.1177/074108839310001002
Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. Discourse Processes, 51(5–6), 511–534. https://doi.org/10.1080/0163853X.2014.910723

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. The Journal of Writing Assessment, 8(1), 1-19.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. Journal of Second Language Writing, 32(1), 1-16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. Journal of Second Language Writing, 18(2), 119-135. https://doi.org/10.1016/j.jslw.2009.02.002

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. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Cruzen, D. (2010). Assessment in the second language writing classroom. University of Michigan.

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

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

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of ‘diagnostic competence’. Language Testing, 21(3), 259–283. https://doi.org/10.1177/0265532204021003001

Erdosy, U. M. (2006). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (Research Report (RR-03-17). Educational Testing Service.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. School Psychology Quarterly, 18(1), 52–65. https://doi.org/10.1521/sqp.181.1.52.20876

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

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metric: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193–202. https://doi.org/10.3758/BF03195564

Guo, L., Crossley, S. A., & McNamara, D. S. (2012). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. Assessing Writing, 18(3), 218-238. https://doi.org/10.1016/j.asw.2013.05.002

Hailliday, M. A. K., & Hasan, R. (2013). Cohesion in English. Routledge.

Hyland, K. (1990). A Genre Description of the Argumentative Essay. RELC Journal, 21(1), 66–78. https://doi.org/10.1177/0336882902003005

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. Journal of Pragmatics, 30(4), 437–455. https://doi.org/10.1016/S0378-2069(97)00009-5

Hyland, K. (2005). Metadiscourse. Continuum. 

Hyland, K. (2008). Second Language Writing. Cambridge University Press.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. Applied Linguistics, 25(2), 156–177. https://doi.org/10.1093/applin/25.2.156

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughley, J. B. (1981). Testing ESL composition: A practical approach. Newbury House Publishers.

Jaliliar, A., & Alipour, M. (2007). How explicit instruction makes a difference: Metadiscourse markers and EFL learners’ reading comprehension skill. Journal of College Reading and Learning, 38(1), 35–52. https://doi.org/10.1080/10790195.2007.10850203

Janssen, T., Vogelin, C., Machts, N., Keller, S. D., & Köller, O. (submitted). Matching the machine: Comparing teachers’ and student teachers’ judgments on student essays to standardized ratings.

Kaiser, J., Südkamp, A., & Möller, J. (2016). The effects of student characteristics on teachers’ judgment accuracy: Differences in gender, ethnicity, minority status, and achievement. Journal of Educational Psychology, 109 (6), 871–888. https://doi.org/10.1037/edu0000156

Keller, S. D. (2013). Integrative Schreibdidaktik Englisch für die Sekundarstufe—Theorie, Prozessgestaltung, Empirie. Narr Verlag.

Knoch, U. (2013). Teacher coherence, considerable strain for reader: A comparison between two rating scales for the assessment of coherence. Assessing Writing, 12(2), 108–128. https://doi.org/10.1016/j.asw.2007.07.002

Knoch, U. (2009). Diagnostic writing assessment—The development and validation of a rating scale. Lang. Knoop, M. Jost, J. Nachtwai, N. Becker-Mrozek, M., & Grabowski, J. (2012). Teilkomponenten von Schreibkompetenz untersuchen: Bericht aus einem interdisziplinären empirischen Projekt. In H. Bayhuber, U. Harms, B. Muszynski, B. Rolle, M. Rothgangel, L.-H. Schön, H. J. Vollmer, & H.-G. Weigand (Eds.), Formate Fachdidaktischer Forschung—Empirische Projekte—Historische Analysen—Theoretische Grundlegungen (Vols. 47–65). Waxmann.

Kruse, O. (2013). Perspectives on academic writing in European higher education: Genres, practices, and competences. Revista de Docencia Universitaria, 11 (1), 37–58. https://doi.org/10.4995/redu.2013.5591

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. TESOL Quarterly, 49(4), 757–786. https://doi.org/10.1002/tesq.194

Lee, I. (2002). Teaching coherence to ESL students: A classroom study. Journal of Second Language Writing, 11(2), 135–159. https://doi.org/10.1016/S1060-3743(02)00065-6

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Testing Language, 19(3), 246–276. https://doi.org/10.1093/2002/02653220123000

Marshall, J. C., & Powers, J. M. (1969). Writing neatness, composition errors, and essay grades. Journal of Educational Measurement, 6(2), 97–101. https://doi.org/10.1111/j.1745-3984.1969.tb00665.x

McCollow, W., & Remstedt, R. (1965). Composition rating scale for general merit: An experimental evaluation. Journal of Educational Research, LIX (2), 55–57. https://doi.org/10.1080/00220671.1965.10883300

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written Communication, 27(1), 57–86. https://doi.org/10.1177/0741088109335154

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metric. Cambridge University Press.
McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(1), 247–288. https://doi.org/10.1080/01638539609544975

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *Educational Testing Service.*

National Assessment Governing Board. (2010). *Writing framework for the 2011 national assessment of education progress.* US Government Printing Office.

National Statistical Office of Switzerland. (2016). Graduation rates – data of figures. Retrieved from https://www.bfs.admin.ch/bfs/de/home/statistiken/berichte-dokumente/unterrichtsmethoden-und-zusammenhaenge-von-analysen-zu-leistungen-einstellungen-unterrichtsmethoden-und-zusammenhaenge-von-leistungen-in-der-mutter-und-fremdsprache.html

Nimechkaïm, V. (2010). Materials for assessing the writing skill. *Advances in Language and Literary Studies, 1*(2), 233–255. https://doi.org/10.7575/aiac.alls.v.1n.1p.233

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

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.* Waxmann.

Rokedzon, 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. https://doi.org/10.1016/j.asw.2016.12.004

Rauin, U., & Meier, U. (2007). Subjective Einschätzungen des Kompetenzerwerbs in der Lehramtsausbildung. In M. Lüders & J. Wissinger (Eds.), *Forschung zur Lehrerbildung–Kompetenzentwicklung und Programmevaluation* (pp. 103–131). Waxmann.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers’ perceptions of young children’s cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360. https://doi.org/10.3102/0027589610374874

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

Ruegg, R., & Sugiyama, Y. (2013). Organization of ideas in writing: What are raters sensitive to? *Language Testing in Asia, 3*(1), 1–13. https://doi.org/10.1116/2329-0643-3-8

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

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. https://doi.org/10.3102/0028312003002125

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. https://doi.org/10.1027/1614-2241/a000016

Schrader, F.-W. (2013). Teacher diagnosis and diagnostic competence. *Beiträge Zur Lehrerbildung, 31*(2), 154–165.

Südkamp, A., & Proterarius, A.-K. (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen.* Waxmann.

Swales, J. (1990). Genre analysis: *English in academic and research settings.* Cambridge University Press.

Tejada, Á., Gallardo, C. N., Ferradá, C. M., & Calderón, L. I. (2015). 2L english texts and cohesion in upper CEFR levels: A corpus-based approach. *Procedia Social and Behavioral Sciences, 212*, 192–197. https://doi.org/10.1016/j.sbspro.2015.11.319

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 33*, 263–271. https://doi.org/10.1037/h0071663

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36*(1), 82–93. https://doi.org/10.1016/j.turtle.2303757609

Vande Kopple, W. J. (1997, March 12). Refining and applying views of metadiscourse [Paper presentation]. Annual Meeting of the Conference on College Composition and Communication, Phoenix, AZ, USA.

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

Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal, 1*, 1–17. https://doi.org/10.1080/09571736.2018.1522662

Weigle, S. C. (2002). Assessing writing. Cambridge University Press.

Weir, C. (1988). The specification, realization and validation of an english language proficiency test. In A. Hughes & T. B. Council (Eds.), *Testing English for University Study. ELT Documents 127* (pp. 45–110). Modern English Publications in association with The British Council.

Witte, S. P., & Faigley, L. (1981). *Coherence, cohesion, and writing quality.* College Composition and Communication, 32*(2), 189–204. https://doi.org/10.2307/356693

Yong, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*(1), 31–48. https://doi.org/10.1016/j.linged.2011.09.004

Zemach, D. E., & Stafford-Yilmaz, L. (2008). Writers at work—the essay. *Cambridge University Press.

Zhu, W. (2001). Performing argumentative writing in English: Difficulties, processes, and strategies. *TESL Canada Journal, 19*(1), 1. https://doi.org/10.18806/tesl.v19i1.918
Appendix A Analytic rating scales for argumentative essays

Organization
4 - Text contains a variety of effective and sophisticated transition words and the organizational structure can be recognized effortlessly
3 - Text contains some effective and sophisticated transition words and the organizational structure can be recognized with little effort
2 - Text contains few effective and mostly simple transition words and the organizational structure can be recognized with some effort
1 - Text contains few and only simple transition words and the organizational structure is difficult to recognize

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 correct spelling and punctuation throughout
3 - Author uses mostly correct spelling and punctuation
2 - Author uses partly correct spelling and punctuation, with some distracting errors
1 - Author uses partly correct spelling and punctuation, with many distracting errors

Grammar
4 - Author uses a variety of complex grammatical structures, few grammar mistakes
3 - Author uses some complex grammatical structures, grammar mostly correct
2 - Author uses few complex grammatical structures, grammar partly correct
1 - Author uses few or no complex grammatical structures, grammar mostly incorrect

Vocabulary
4 - Author uses sophisticated, varied vocabulary throughout
3 - Author mostly uses sophisticated, varied vocabulary
2 - Author partly uses sophisticated, varied vocabulary, sometimes repetitive
1 - Author uses little sophisticated, varied vocabulary, often repetitive

Overall task completion
4 - Text fully addresses the essay question
3 - Text mostly addresses the essay question
2 - Text partly addresses the essay question
1 - Text does not address the essay question
Appendix B Holistic scale: NAEP Levels (National Assessment of Educational Progress)

Score = 6 Effective skill in responding to the writing task.
- The response formulates a clear position that fully addresses the complexities of the issue.
- The response provides strong persuasive reasons and evidence to support the writer’s position.
- Organization demonstrates a logical, well-executed progression of ideas that effectively supports the persuasive purpose. Transitions effectively convey relationships among ideas.
- Sentence structure is varied to communicate relationships among ideas. Word choice is precise and supports the persuasive purpose.
- Though there are a few minor errors in grammar, usage, and mechanics, meaning is clear throughout the response.

Score = 5 Competent skill in responding to the writing task.
- The response formulates a position that usually recognizes multiple significant aspects of the issue but may not fully address some of its complexities.
- The response usually provides persuasive reasons and evidence to support the writer’s position.
- Organization is clear and demonstrates a logical progression of ideas that supports the persuasive purpose. Transitions usually convey relationships among ideas.
- Sentence structure usually communicates relationships among ideas. Word choice is usually precise.
- Grammar, usage, and mechanics are usually correct with a few distracting errors, but meaning is clear.

Score = 4 Adequate skill in responding to the writing task.
- The response takes a position and acknowledges significant aspects of the issue.
- While details and examples provide adequate evidence to support the writer’s position, their development is somewhat uneven.
- Ideas are focused on the topic and adequately reflect the writer’s argument.
- An organizational structure is evident and relationships among ideas are mostly clear.
- Sentence structure is somewhat varied to communicate relationships among ideas. Word choice is mostly clear.
- Grammar, usage, and mechanics are mostly correct, with some distracting errors.

Score = 3 Developing skill in responding to the writing task.
- The response states a position but addresses only some of the aspects of the issue. It shows little understanding of other perspectives.
- Some relevant reasons and evidence for the writer’s position are used, but they are unevenly developed.
- The response uses a simple organizational structure, and relationships among ideas are sometimes unclear.
- There is a little sentence variety to communicate relationships among ideas. Word choice at times is not appropriate for the writer’s purpose.
- Grammar, usage, and mechanics show some distracting errors that impede understanding.

Score = 2 Marginal skill in responding to the writing task.
- The response does not state a clear position and provides few reasons to support the writer’s view.
- Relationships among ideas are often illogical or unclear.
- There is little, if any, sentence variety. Word choice is rarely specific and does little to support the persuasive purpose.
- Grammar, usage, and mechanics show many distracting errors that impede understanding.

Score = 1 Little or no skill in responding to the writing task.
- The response gives a few simplistic reasons. If any details or examples are present, they are brief, general, undeveloped, or not relevant to the persuasive purpose.
- Groupings of ideas are mostly illogical.
- Sentence structure is incorrect; word choice is unclear and inappropriate.
- Grammar, usage, and mechanics frequently contain distracting errors that impede understanding.
