frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/feduc.2022.784261

Assessing the Quality of Argumentative Texts: Examining the General Agreement Between Different Rating Procedures and Exploring Inferences of (Dis)agreement Cases

Yana Landrieu, Fien De Smedt, Hilde Van Keer and Bram De Wever*

Department of Educational Studies, Ghent University, Ghent, Belgium

OPEN ACCESS

Edited by:
Marije Lesterhuis, Spaarne Gasthuis, Netherlands

Reviewed by:
Stefan Daniel Keller, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
Gustaf Bernhard Uno Skar, Norwegian University of Science and Technology, Norway

****Correspondence:***
Yana Landrieu
Yana.Landrieu@UGent.be

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *27 September 2021*
Accepted: *29 March 2022*
Published: *04 May 2022*

Citation:
Landrieu Y, De Smedt F, Van Keer H and De Wever B (2022) Assessing the Quality of Argumentative Texts: Examining the General Agreement Between Different Rating Procedures and Exploring Inferences of (Dis)agreement Cases. Front. Educ. 7:784261. doi: 10.3389/feduc.2022.784261

Assessing argumentative writing skills is not a straightforward task, as multiple elements need to be considered. In function of providing feedback to students and keeping track of their progress, evaluating argumentative texts in a suitable, valid and efficient way is important. In this state-of-the-art exploratory study, 130 argumentative texts written by eleventh graders were assessed by means of three different rating procedures (i.e., absolute holistic rating, comparative holistic rating, and absolute analytic rating). The main aim of this study is twofold. First, we aim to examine the correlations between the three rating procedures and to study the extent to which these procedures differ in assigning scores. In doing so, the more innovative approach of pairwise comparisons is compared to more established assessment methods of absolute holistic and analytic rating. Second, we aim to identify key characteristics that determine the quality of an argumentative text, independent of the rating procedure used. Furthermore, key elements of mid-range, weak and strong argumentative texts were studied in detail. The results reveal low to moderate agreement between the different procedures, indicating that all procedures are suitable to assess the quality of an argumentative text; each procedure, however, has its own qualities and applicability.

Keywords: argumentative writing, rating procedures, holistic rating, analytic rating, pairwise comparisons

INTRODUCTION

Effective writing skills are considered imperative in our twenty-first century society, as they are highly valued in private, educational, and professional contexts (Graham and Perin, 2007). This is especially true for argumentative writing skills. Argumentative writing skills are considered important as they help to clarify our thoughts and make us reflect on the thoughts of others (by integrating different points of view) and stimulate critical thinking and problem-solving competences (Varghese and Abraham, 1998; Nussbaum and Schraw, 2007; Granado-Peinado et al., 2019). However, the majority of students experience difficulties developing effective writing skills in general, and more particularly in the genre of argumentative writing (NCES, 2012). The

argumentative writing proficiency of students appears to be highly substandard (Graham and Perin, 2007; NCES, 2012; Ferretti and Lewis, 2013; Song and Ferretti, 2013; Traga Philippakos and MacArthur, 2019). Ferretti and Lewis (2013), for example, found that students' argumentative texts rarely acknowledge opposing positions, rarely consider the merits of different views, and almost never include rebuttals of alternative perspectives.

Determining the quality of an argumentative text is not a straightforward task as different elements need to be considered. Nevertheless, with regard to providing feedback to students, keeping track of their progress, and helping them to write better texts, it is important to be able to evaluate argumentative texts in a suitable, valid, and efficient way. By taking a closer look at the texts in our sample, we have gained insights regarding features of stronger and weaker argumentative texts, which will be shared in this study. In what follows, we firstly present three rating procedures that are central in this study: (1) Absolute holistic rating, (2) comparative holistic rating, and (3) absolute analytic rating. As comparative holistic rating is an innovative and upcoming assessment procedure in writing research, we will compare this rating procedure to more established methods such as absolute holistic rating and absolute analytic rating. Next, we briefly review the literature on the assessment of argumentative texts. More specifically, we discuss the need to assess (1) the quality of argumentation, (2) the quality of content, and (3) the inclusion of general text characteristics to determine the overall quality of an argumentative text. The main aim of this exploratory study is to compare (a) different rating procedures that can be used when assessing argumentative texts, and (b) to identify text features of weak and strong argumentative texts. This study is innovative as this is the first study comparing the three rating procedures, especially given that pairwise comparisons are not yet as widespread and established as holistic and analytic rating. Secondly, we closely examine the specific features of a weak or strong argumentative text. Which features make a text weak or strong? Combining these two insights can be informative for assessment practices and give more insight into the key aspects of an argumentative text, regardless of the rating procedure used.

THEORETICAL BACKGROUND

Rating Procedures to Assess Text Quality

It is essential to assess the quality of argumentative texts in a suitable and valid way. Selecting a rating procedure is, however, not easily decided. In determining the most suitable procedure, a number of factors, such as the available time, the aim of the assessment, and the amount of raters and texts, should be taken into account. A review of previous research shows that many different procedures are used to assess written texts. These procedures differ in the degrees of rating freedom. Following Coertjens et al. (2017), rating procedures can be classified in two dimensions: Holistic vs. analytic on the one hand and absolute vs. comparative on the other hand (see also Harsch and Martin, 2013; Bouwer and Koster, 2016; Coertjens et al., 2017). In holistic rating, texts are rated as a whole, whereas in analytic rating, text quality is measured by scoring multiple features of a text. In absolute ratings, every text is scored by a description or a criteria list, whereas in comparative ratings, texts are compared to each other to assess the text quality. In this study, we focus on three rating procedures: (1) Absolute holistic rating, (2) comparative holistic rating, and (3) absolute analytic rating.

Absolute Holistic Rating

Within absolute holistic rating, there are differences regarding the extent to which a rater has access to specific rating criteria. For instance, a holistic rubric provides the rater with predefined rating criteria. In this way, raters using such rubrics still provide a holistic assessment based on their overall impression of a text but they are supported by the holistic explanations provided with each text score (Penny et al., 2000; Yune et al., 2018).

Another way to holistically assess a text is general impression marking. Following this procedure, texts are rated as a whole by assigning a score based on a total impression (Charney, 1984). Raters receive a general description regarding the assignment and the competences that are being pursued while writing. However, raters do not receive explicit rating criteria to assign a particular score. Each rater has (unconsciously) an internal standard on how to evaluate a text, *inter alia*, based on earlier rating experiences. An advantage reported in the literature is that this procedure does not require a lot of time and effort, as scores are rather quickly assigned without explicit rating criteria (Charney, 1984).

There are two drawbacks linked to general impression marking: rater variance and the lack of detailed feedback on students' performance (Carr, 2000; Weigle, 2002; Lee et al., 2009). Regarding rater variance, not every rater uses the full scale to assign scores. For instance, raters can vary in terms of rigor by systematically assigning either higher or lower scores to texts. Additionally, raters can also have different rating criteria in mind or can perceive some elements as more important than other elements, even though they are asked to rate holistically (Weigle, 2002; Lee et al., 2009; Bouwer and Koster, 2016). Another explanation for varying scores is the *halo effect*, as described by Thorndike (1920). "Ratings *are* apparently affected by a marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of the qualities by this general feeling" (p. 25). The quality of general impression marking may also depend heavily on the experience of the assessors. Rater training and experience could increase the reliability between raters, but this is not automatically the case (Myers, 1980; Charney, 1984; Huot, 1993; Rezaei and Lovorn, 2010; Coertjens et al., 2017). To reduce rater variance, support (i.e., rater training, or support in using the whole scale) for holistic raters is thus essential. By doing so, the reliability of the ratings can be increased (Bouwer and Koster, 2016). However, when raters are supported with criteria, we no longer apply general impression marking, as this is a rating procedure that works without rating criteria. As to the second drawback, general impression marking does not provide insight into students' weaknesses and strengths in (argumentative) writing in detail. In this respect, a general score is assigned to a text without providing any detail or information on how and why this particular score was assigned.

Nevertheless, teachers can provide additional feedback so the student has insight into the strengths and weaknesses of the text.

Whenever absolute holistic rating is mentioned in this study, we are referring to general impression marking. We chose to implement absolute holistic rating in this way, as many teachers in schools still use this approach when evaluating argumentative texts and we were also able to observe this in Flanders.

Comparative Holistic Rating (Pairwise Comparisons)

In comparative holistic rating, texts are holistically compared to each other to assess the text quality. A well-known comparative holistic rating approach is pairwise comparison. The holistic character implies that raters are free to define how to assess the texts without any predetermined criteria (van Daal et al., 2016). The comparative character implies that each rater compares two texts and selects the best one. This is applied multiple times and creates a binary decision matrix of the worst and the best text in each comparison (Coertjens et al., 2017). This results in a reliable ranking order of texts ranging from the worst rated text to the best rated text. Texts are constantly compared to each other, and each text is evaluated multiple times, by multiple raters. This procedure is based on Thurstone's law of comparative judgment (1927) which explains how objects (e.g., written argumentative texts) can be scaled from lowest to highest text quality by pairwise comparisons (Pollitt, 2012). Following Thurstone (1927), raters are more competent in comparing two different texts to each other than to rate one text as a whole (Thurstone, 1927; Gill and Bramley, 2013; McMahon and Jones, 2015). Multiple raters compare two different texts and select the best one, according to their opinion. By using the Bradley-Terry model, a scale from worst to best text can be generated (McMahon and Jones, 2015; Coertjens et al., 2017). By using this scale, teachers or writing researchers can easily assign a score to a text. This method originated in psychophysical research but has become applicable for educational assessment purposes as well (Pollitt, 2012; McMahon and Jones, 2015).

This procedure is easy to implement, as raters simply have to decide which of the two presented texts is the best one (McMahon and Jones, 2015). By doing so, pairwise comparisons eliminate differences in the severity of raters (van Rijt et al., 2021). Another advantage is that there is no need for an extensive training procedure for raters. However, deciding which text is better can be difficult in some instances (e.g., a text with high-quality content, but with poor argumentative structure or two texts of a similar level). Therefore, raters need a clear understanding of the writing assignment goals to assess which text is the best one. Pairwise comparisons are not easily applicable in regular teaching activities, as multiple raters are required to achieve a reliable scale. Due to the fact that this procedure can be difficult to implement in a school context, this procedure is sometimes considered inefficient (Bramley et al., 1998; Verhavert et al., 2018).

Overall, the reliability of pairwise comparisons appears to be much higher compared to absolute holistic rating procedures (Thurstone, 1927; Pollitt, 2012; Gill and Bramley, 2013). The reliability of pairwise comparisons depends on the amount of comparisons: The more comparisons, the more reliable the ranking order (Bouwer et al., in review)[1]. Next to a high reliability, pairwise comparisons also provide valid scores (Pollitt, 2012; van Daal et al., 2016). Each text is evaluated by multiple raters and the final ranking order is a reflection of the multiple raters' expertise (Pollitt, 2012; van Daal et al., 2016). This implies that pairwise comparisons result in a reliable ranking order. Individual rater effects can be neglected, due to the large number of raters, which ensures that each text can be compared several times with another text (e.g., in this study each text is, on average, compared 16.6 times to another text).

Pairwise comparisons are not the only way to assess texts in a comparative, holistic way. Benchmark rating could also be a way of comparatively and holistically assessing an argumentative text, by providing raters benchmarks that each represent a certain text quality. For more information on this comparative holistic rating procedure, we refer the reader to Bouwer et al. (see text footnote 1). As we opted to use pairwise comparisons in this study, benchmark rating will not be further explored.

Absolute Analytic Rating

Analytic rating procedures are more detailed than holistic procedures, as text quality is measured by scoring multiple features of a text (i.e., sub scores for specific text features or facets that a rater has to keep in mind) which can be added up (Harsch and Martin, 2013; Coertjens et al., 2017). There are several advantages linked to this procedure. First, by using an analytic rating procedure, weaknesses and/or strengths in a text can be distinguished, leading to more information for teachers or researchers. This can lead to more precise feedback which can improve the learning process of the student (Lee et al., 2009). Second, earlier research (Vögelin et al., 2019) showed that lexical features can have an influence on how text quality is rated; however, the chance that one specific weakness in a text (e.g., grammar) is decisive in the overall assessment is smaller for analytic procedures than it is with holistic rating (Barkaoui, 2011). Third, by defining rating criteria in advance, more equal and reliable scores between raters can be obtained. Previous research on reliability of analytic rating is, however, still very inconsistent. Earlier research of Follman et al. (1967) and Charney (1984) claims that absolute analytic rating does lead to good or increased reliability compared to absolute holistic rating, whereas research by Goulden (1994) and Barkaoui (2011) claims that analytic rating leads to decreased reliability. In addition, training the raters could increase reliability and validity, but this does not automatically lead to reliable and valid scores (Rezaei and Lovorn, 2010; Harsch and Martin, 2013). Therefore, Harsch and Martin (2013) suggest combining holistic and analytic rating procedures to achieve more reliable and valid results.

In addition to the enumerated benefits, previous research also reported several drawbacks related to analytic rating. More specifically, analytic rating can be time consuming, as each text feature is separately scored (Hunter, 1996). In this respect, it is questionable whether the sum of the parts is a representative score of a text (Huot, 1990). When writing researchers or teachers want to achieve a reliable score, multiple raters can be used. This,

[1]Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., and De Maeyer, S. (in review). *Comparative Approaches to the Assessment of Writing: Reliability and Validity of Benchmark Rating and Comparative Judgement*.

however, makes it more difficult to apply in practice (Lee et al., 2009). During the assessment, raters may not be able to call upon their expertise (and do not have ownership of the total score), as they are tied to rating the predetermined criteria (in contrast to absolute and comparative holistic rating). Additionally, analytic rating does not automatically create rich data, as all elements are simply added up (Hunter, 1996). In other words: Analytic rating does not look at the whole picture, as opposed to absolute and comparative holistic rating. Time must be invested in setting up an analytical rating procedure.

A frequently used analytic rating procedure is the use of rubrics. By using an analytic rubric, written texts are rated on multiple aspects and sub scores are allocated considering specific text features or facets that a rater has to keep in mind (Weigle, 2002; Barkaoui, 2011; Harsch and Martin, 2013). By adding up the sub scores, an overall score can be assigned. The goal of using the rubric-criteria is to enlarge the agreement between different assessors and thus reduce rater variability. In an analytic rubric, the text features are predetermined, but the weight of these text features is not always determined in advance. Following Sasaki and Hirose (1999) and Coertjens et al. (2017), raters can independently decide which weight they give to the text features. This implies that a text feature to which the rater attaches great importance can be more decisive than another text feature. Other authors, like Stapleton and Wu (2015), describe the weight of the separate text features in a rubric as fixed. This implies that the rater cannot decide the weight of each text feature and this makes analytic rating less free than holistic rating.

Determining the Overall Quality of an Argumentative Text

Quality of Argumentation

Toulmin (1958) argued that an argumentative text is composed of (a) a claim, (b) data, (c) warrants, (d) backings, (e) qualifiers and (f) rebuttals. The claim is the thesis of the author, whereas data is the foundation for the claim. A warrant is the relation between the data and the claim. Backings are additional evidence that support the claim. A qualifier adds credibility to the argument, whereas rebuttals are circumstances under which a claim is not valid (Toulmin, 1958). The original Toulmin-model has been modified in contemporary literature into a more understandable and practical model (Nussbaum and Kardash, 2005; Nussbaum and Schraw, 2007; Qin and Karabacak, 2010; Stapleton and Wu, 2015). Alongside the work of Toulmin (1958) and Stapleton and Wu (2015) stated that a strong argumentative text is composed of two important elements. First, an argumentative text must be constructed taking into account all elements contributing to a good *quality of argumentation*. Second, attention must be paid to the *quality of the content* of the text. According to Clark and Sampson (2007) and Stapleton and Wu (2015), many studies prioritize the importance of the quality of argumentation over the quality of content. As Clark and Sampson (2007) mention, the majority of research on argumentative writing skills focuses explicitly on the Toulmin-structure, without paying attention to the content of the argumentative text leading to an incomplete picture of the quality of the text (Simon, 2008). In line with Stapleton and Wu (2015), we therefore argue that it is not self-evident, but important to take both the quality of argumentation, the quality of the content and the general text characteristics into account when evaluating an argumentative text, as all three elements are connected and cannot be completely separated.

Quality of Content

In addition to the quality of argumentation, previous studies also examined the quality of the content. In this respect, three criteria are distinguished in the literature: overall persuasiveness, factual accuracy, and information originating from source texts. First, as the main goal of argumentation is to convince or persuade an audience of a certain point of view, a high-quality argumentative text should have a good overall persuasiveness (De La Paz and Felton, 2010). Strong persuasive arguments require deep reasoning from students, as they need to come up with good reasons to support the claim (Marttunen et al., 2005). Second, an argumentative text should be factually accurate (De La Paz and Felton, 2010). Third, the author should integrate information originating from multiple, reliable source texts into one's argumentative text (De La Paz and Felton, 2010; Cuevas et al., 2016). This implies that the author needs to consider the multiple points of view that are present in the source texts (Wolfe and Britt, 2008). Writers must have the capacity to draw upon evidence to support their point of view (Kibler and Hardigree, 2017). It is allowed for writers to express their own opinions, but it is recommended that they support these opinions with objective sources.

General Text Characteristics

As well as considering the quality of argumentation and the quality of content, various general text characteristics also appear to be key in determining the overall quality of an argumentative text. More particularly, including an *introduction* and/or *conclusion* in a text can be helpful for the reader. A good introduction draws the reader's attention and reveals the main topic of the text to the reader, and by reading the conclusion, readers can quickly find out the point of view of the author (Syed et al., 2021).

In addition, as Barkaoui (2010) and Wolfe et al. (2016) mention, *text length* significantly influences text quality. Longer texts contain more information and details and are therefore often associated with a higher text quality. However, including unnecessary and *irrelevant information* in texts can hinder the flow and readability of a text. Finally, *bad writing mechanics* seem to negatively affect text quality (Figueredo and Varnhagen, 2005; Rezaei and Lovorn, 2010; Jansen et al., 2021). However, this list is not exhaustive. There are many other elements (e.g., structure, logical line of reasoning, etc.) that determine text quality, but these are out of the scope of this study.

The Present Study

A variety of assessment methods exists, but the literature generally distinguishes between holistic and analytic rating procedures, as discussed in the theoretical background. There appears to be a misconception that the use of analytic rating automatically leads to a reliable score. As the results in the educational field seem to be inconsistent and reveal mixed results

on reliability and validity (e.g., Charney, 1984; Barkaoui, 2011), more research is needed. Harsch and Martin (2013) reveal that both holistic and analytic rating procedures have their strengths and weaknesses, depending on the purpose for which they are used (see also Barkaoui, 2011). More recently, another distinction in rating procedures has been identified in the literature: absolute and comparative rating procedures (Coertjens et al., 2017). In this respect, a comparative approach by means of pairwise comparisons has been introduced to effectively and efficiently assess students' writing performance (Coertjens et al., 2017). Pairwise comparisons are proven to be a valid and reliable rating procedure and therefore seem to be a promising alternative for absolute holistic and absolute analytic rating procedures (van Daal et al., 2016; Coertjens et al., 2017). To date, there is no research yet that focuses on comparing these three rating procedures. Therefore, this study will tackle this issue. The main aim of this study is twofold. The first aim of this study is to examine the correlations between the three rating procedures and to study the extent to which these procedures differ in assigning scores. In doing so, the innovative approach of pairwise comparisons is compared to more established assessment methods of absolute holistic and analytic rating.

In this study, we choose to use pairwise comparisons as a starting point for describing results which we then use to make connections to the other rating procedures. There are three reasons for this approach. First, pairwise comparisons use multiple raters, leading not only to a high level of reliability, but also to a broadly based consensus. Research of Verhavert et al. (2018) showed that the Scale Separation Reliability (SSR) can be interpreted as an inter-rater correlation. Second, whereas holistic and analytic rating are more established and more often used in practice, pairwise comparisons are already commonly used in educational research and are considered promising methods to assess writing performance (Coertjens et al., 2017; Verhavert et al., 2018). The rating procedure is easy to implement for researchers, as specific software exists and raters do not need a lot of training, and it provides opportunities to achieve high inter-rater reliability. However, it also requires a lot of different raters, so this rating procedure is less suitable to use in daily practice. Third, in this study, the use of pairwise comparisons is a procedure that takes into account both quality of content and quality of argumentation. These three arguments ensure that this rating procedure is an optimal procedure to start from and to compare to the other two rating procedures.

The second aim of this study is to identify key characteristics that determine the quality of an argumentative text, independent of the rating procedure used. Regarding the second aim, in addition to making an informed choice regarding the assessment procedure, the evaluator must also have an understanding of the essential criteria of an argumentative text. Based on previous research by Stapleton and Wu (2015), the overall quality of an argumentative text is determined by the quality of argumentation and the quality of content. In addition, several general text characteristics (e.g., the inclusion of an introduction and conclusion, text length, use of irrelevant information and writing mechanics) should be taken into account as they influence (argumentative) text quality as well. Therefore, we want to identify key characteristics that determine the quality of an argumentative text. In this respect, we particularly focus on examining the elements that seem to be associated with mid-range, weak and strong argumentative texts. Based on the twofold aim of the study, three main research questions are addressed in the present study.

RQ1a: How do absolute holistic rating, comparative holistic rating (pairwise comparisons) and absolute analytic rating correlate?

RQ1b: How often do we see deviations between these rating procedures and how strong are these deviations?

RQ2: Which elements characterize mid-range, weak and strong argumentative texts, independent of the rating procedure used?

MATERIALS AND METHODS

Participants

In total, 164 eleventh grade students participated in the study and wrote an argumentative text. Students were on average 17 years old, their age varying between 16 and 19 years. All students were enrolled in the academic track of secondary education. The majority of the students were native Dutch speakers (n = 156, 95.1%), 3.7% were bilingual (Dutch + another home language) (n = 6) and 1.2% had another home language (n = 2) (French). The majority of the participants were female (n = 123, 75%).

Data Collection Procedure

After signing an active informed consent (the parents/guardians received a passive informed consent), the students had to complete an argumentative writing test. Half of them (n = 79) completed a digital writing test on the conservation of zoos, and the other half completed a digital writing test on voting rights from the age of 16 (n = 85). Each student received two source texts on the respective topic and was instructed to write an argumentative text based on the source texts and based on their own opinion. This integrated writing task required the secondary school students to write an argumentative text (with the goal to persuade the reader) by using the informative source texts. They were free to choose their own point of view and (counter) arguments and rebuttals. The secondary school students were not allowed to copy-and-paste from the source texts, but they were asked to integrate the arguments from the informative source texts into their own argumentative texts (in their own words). They were free to add additional arguments or other information, not directly drawn from the source texts. They were allowed to use a digital draft sheet, but were not allowed to search for extra information on the internet. The source texts were similar in difficulty and length (i.e., on average 634 words). Furthermore, the students were instructed to clearly take a stand and defend one position. They had to write individually and had to complete the argumentative writing test within 45 min, without further guidance.

Due to the Covid-19 pandemic, the data collection was discontinued abruptly. Nevertheless, we were able to collect 164 texts in total, of which 157 texts were further included in the study (i.e., due to late submission, seven texts could not be assessed using the three rating procedures). Although the assignment

explicitly stated to write an argumentative text, 27 texts did not take a position (e.g., pro or contra), did not have the goal to persuade, nor were any arguments integrated. Therefore, these texts were categorized as informative and eliminated from further analyses. 130 argumentative texts with an average length of 401 words (SD = 113, min = 166, max = 873) were included in the analyses. All texts were anonymized. Raters were unaware of the gender and language background of the authors of the texts.

Rater Training and Rater Procedures

Raters

In light of a research assignment on assessment, university students enrolled in the second year of educational sciences (n = 132) collected the data. Prior to the data collection, the definition and the goal of argumentative writing were explained to the university students, and they were introduced to the differences between rating procedures. Furthermore, they received a protocol outlining the data collection procedure, which they had to follow strictly. After collecting the data, these 132 university students also served as raters for the pairwise comparisons. The holistic and analytic rating procedure were executed by the researcher and a trained rater (n = 2) (see **Table 1**).

Instructions for Holistic and Analytic Rating

The argumentative texts were holistically and analytically rated by the first author and a trained language teacher who teaches Dutch in secondary education (see **Table 1**). According to Bacha (2001), training additional raters in how to assess texts is key. Therefore, the second rater received an instruction guide and followed an intensive training session given by the first author (3–4 h). During this session, the structure of an argumentative text was explained in detail. More particularly, the adapted model of Toulmin, as used by Nussbaum and Kardash (2005) and Stapleton and Wu (2015), was instructed and each element of the model was illustrated by means of specific examples. Furthermore, both holistic and analytic rating were explained in detail and the specific assessment procedures were discussed and practiced. During practice, ten texts (on both writing topics) were rated holistically and analytically and discussed with the first author.

For holistic rating, no specific instructions were given to the rater except for the instruction to assign a holistic score from 0 to 10 that best reflects the quality of this argumentative text. The goal was to intuitively map the quality of the text according to a general impression without predefined criteria, as Myers (1980) recommends.

For analytical rating, the raters used the framework developed by Stapleton and Wu (2015), the so-called "Analytic Scoring Rubric for Argumentative Writing" (ASRAW). In the ASRAW, quality of argumentation is determined by looking at six elements, based on the earlier research of Nussbaum and Kardash (2005) and Qin and Karabacak (2010). The elements are: (a) A claim, (b) claim data, (c) a counterclaim, (d) counterclaim data (e), rebuttals, and (f) rebuttal data. **Table 2** provides an overview of these elements, including a description for each element. Ideally, all elements are included in a logically structured argumentative text. So the more a text conforms to the (adapted) Toulmin-structure, the stronger and more persuasive it can be (Qin and Karabacak, 2010). However, when a text does not include all elements, the text is not automatically considered a weak text. Much also depends on the quality of the content and the general text characteristics (Stapleton and Wu, 2015). The order in which the elements appear is neither linear nor predetermined (e.g., a text does not have to start with a claim, the counterclaim and counterarguments can be placed before the actual claim).

The ASRAW uses different performance levels (for claim data, counterargument data and rebuttal data) and a dichotomous scale (for claims, counterargument claims and rebuttals). Each rating dimension is given a score, and although the weight of the elements is predetermined, not all elements are given the same weight (e.g., if a text mentions a claim, a score of 5 is given; if a text mentions a counterargument claim, a score of 10 is given). The specific weight attached to each element was decided by Stapleton and Wu (2015), the original developers of the framework. As data, counterarguments, and rebuttals require a higher level of critical thinking and argumentation skills, a higher weight is given to these elements. By adding up the scores, a total score is presented for the whole argumentative text. Scores ranged from 5 to 100. For more detailed information, we refer to Table 4 in the original work from Stapleton and Wu (2015). As mentioned in the literature overview, the ASRAW seems to prioritize quality of argumentation over quality of content. For instance, a text that does not provide any data (i.e., arguments that defend the point of view) is automatically assigned score "0" for that element, whereas the content of the text might be good. Without a solid argumentative structure, an argumentative text can never receive a high final score according to the ASRAW.

Instructions for Pairwise Comparisons

Argumentative texts were assessed through pairwise comparisons by 132 university students (see **Table 1**). The platform Comproved (Comproved.com) was used to make the comparisons. Pollitt (2012) argues that raters do not need much training when comparing texts to each other (see also Coertjens et al., 2017). Therefore, only a few instructions were given to the raters. The instructions were: "Hen judging which argumentative text is the best one, you can keep the following criteria in mind: (1) The author takes a reasoned position, (2) the author substantiates the position with relevant arguments, (3) the author uses information from sources or presents their own reasoning to support their position, and (4) the text is comprehensible (cf., coherent text structure, sentence structure and word choice)." Correct spelling, use of punctuation and capitalization were not taken into account in the assessment. Alongside these instructions, the raters were also instructed on the genre of an argumentative text by providing them with a definition of argumentative writing and explaining the goal of this genre (i.e., persuading). Given the holistic and comparative nature of this assessment, we did not provide further explicit instruction on the different elements of strong argumentative texts to avoid raters checking for each Toulmin-element in an analytic way.

TABLE 1 | Overview of the used rating procedures, the amount of raters, and the assessment methods.

Rating procedure	Amount of raters	Assessment method
Holistic rating	*n* = 2 (The researcher and a trained rater)	General impression marking, without predefined criteria
Analytic rating	*n* = 2 (The researcher and a trained rater)	By the use of the ASRAW (Stapleton and Wu, 2015)
Pairwise comparisons	*n* = 132 (132 second year educational sciences students)	By the use of the platform Comproved (Comproved.com)

During the rating process, raters were shown two texts each time and they had to select which one was the best argumentative text. Each text was rated multiple times, by multiple raters. More particularly, each student rated 20 pairs of texts independently at home and each text was compared on average 16.6 times to another text. The informative texts (i.e., texts missing a position and arguments) were then eliminated from the data and a ranking from the worst rated text to best rated text was calculated.

Procedures to Obtain Inter-Rater Reliability

Holistic and Analytical Rating

After the training, both the first author and the second rater assessed texts individually and independently. The assessment followed a two-stage process. During the first stage, all texts were rated holistically. During the second stage, texts were rated analytically but in a different order and with 1 week in between to avoid dependency between the two procedures.

For both holistic and analytic rating, the first author rated all argumentative texts (n = 130 texts) and the second rater double coded 24% (n = 31) of the texts. The Intraclass Correlation Coefficient (ICC) of the holistic and analytic ratings was examined based on the two-way mixed model, measuring consistency between raters. For analytic rating, the ICC of the total score of the ASRAW was 0.98, while the ICC for holistic rating was 0.48 (for more detailed per rating procedure information, see **Table 3**).

There are large discrepancies between the ICCs of the holistic and analytic rating procedure. The analytic rating procedure (the ASRAW) appears to be a reliable way to assign scores to argumentative texts. The units of analyses were indicated in advance, which made it easier and more transparent for the rater to assign subscores (as each unit represents an element of an argumentative text), which could partially explain the high ICC. In the holistic rating procedure (general impression marking), we observe a low ICC of 0.48, which is in line with our predictions, as this is an intuitive score, assigned without predefined criteria.

TABLE 2 | Overview of the elements of an argumentative text with a definition of each element, based on Stapleton and Wu (2015).

Elements of an argumentative text	Definition
Claim	An assertion or opinion to a specific topic
Claim data	Data that supports the actual claim
Counterclaim	The possible opposing views contrary to the own claim
Counterclaim data	Data that supports the counterclaim
Rebuttal	A claim that refutes the counterclaim, by responding to the counterclaim
Rebuttal data	Evidence to support the rebuttal

TABLE 3 | Reliability measures per rating procedure.

Collected texts	
Holistic rating	ICC = 0.48
Analytic rating	ICC of the total score of the ASRAW = 0.98 ICC of the individual elements of the ASRAW: • Claim: ICC = 1 • Claim data: ICC = 0.91 • Counterargument: ICC = 0.85 • Counterargument data: ICC = 0.98 • Rebuttal: ICC = 1 • Rebuttal data: ICC = 0.95
Pairwise comparisons	SSR = 0.83

Pairwise Comparisons

After all texts were rated in Comproved, a rank order was generated ranging from the lowest to the highest text quality. In this way, a logit score for each text was estimated. The higher the logit score, the better the text. Research by Verhavert et al. (2018) states that Separation Scale Reliability (SSR) is a good way to check the inter-rater reliability as it can estimate the level of agreement between the multiple raters. SSR is derived from Rasch modeling and is, according to Verhavert et al. (2018), typically used as a reliability measure. SSR is comparable to the ICC for multiple raters, both reflecting reliability of average scores across raters (Verhavert et al., 2018; see text footnote 1). An SSR of 0.80 and higher indicates a high inter-rater reliability. In this study we obtained an SSR of 0.83 (see **Table 3**).

Data Analysis

Preparatory Analyses

Given that the majority of our participants were female and native Dutch speakers, preparatory analyses were conducted to study the relationship between home language and text quality on the one hand, and the relationship between gender and text quality on the other hand. Based on ANOVA analyses, results showed no significant relationships between home language and text quality [pairwise comparisons: $F_{(1,\ 156)} = 0.09$, $p = 0.76$; holistic: $F_{(1,\ 162)} = 0.33$, $p = 0.57$; and analytic: $F_{(1,\ 162)} = 0.31$, $p = 0.58$]

TABLE 4 | Scoring "writing mechanics" of an argumentative text.

	Score 2	Score 1	Score 0
Writing mechanics	> 2 Spelling errors and >2 Syntax errors	1–2 spelling errors or/and 1–2 syntax errors	No spelling errors and No syntax errors

nor between gender and text quality [pairwise comparisons: $F_{(1, 156)} = 0.82$, $p = 0.37$; holistic: $F_{(1, 162)} = 0.31$, $p = 0.58$; and analytic: $F_{(1, 162)} = 0.46$, $p = 0.50$]. In addition, given that text length and writing mechanics are key predictors of text quality, both variables were taken into account in the analyses. For text length, the number of words were counted. For writing mechanics, the following scoring was applied (see **Table 4**). Both text length and writing mechanics were not double coded, as evaluating them was not ambiguous.

The relationships between text length and text quality (for each rating procedure) on the one hand, and writing mechanics and text quality (for each rating procedure) on the other hand, were all significant except for the relation between text length and analytical text quality. Variance explained by (1) text length, (2) writing mechanics, and (3) a combination of both was, respectively, 28.9, 5.5, and 37.5% for pairwise comparisons, 2.6, 3.1, and 6.4% for the analytic rating procedure, and 8.7, 5.54, and 15.8% for the holistic rating procedure.

Furthermore, results of the preparatory analyses showed that the explained variance of text length for pairwise comparisons (28.9%) was the highest and quite substantial. A possible explanation might be that pairwise comparisons are more prone and sensitive to text length, as longer texts were often rated as more qualitative and better texts. In educational research, text length has often been proven to have a significant relationship with text quality (Jarvis et al., 2003; Lee et al., 2009). However, when comparing texts to each other (like pairwise comparisons do), text length is an easy criterion to use. After all, this is the first visual indicator you see when you are presented with text A and text B. With the absolute rating procedures (absolute holistic and absolute analytical rating) you do not have this foundation of comparison. In addition, absolute analytic scoring (the ASRAW) may be less prone to this, due to the specific criteria that are not focusing on text length.

Main Analyses

Main Analyses in View of RQ1a + RQ1b

To study RQ1a and RQ1b, general analyses were conducted on all 130 argumentative texts. To analyze the results, correlations and attenuated correlations were calculated for RQ1a. Concerning the correlation between rating procedures, Bouwer and Koster (2016) stated that: "Since the rating procedures will not have a perfect reliability due to measurement error, correlations between scores from two rating procedures will suffer from attenuation." (p. 43). Therefore, we conducted corrections on the correlations to deal with unreliability and to reflect the true correlations between rating procedures (Bouwer and Koster, 2016). More specifically, we divided the observed correlation coefficient by the product of the square roots of the two relevant reliability coefficients (Lord and Novick, 1968; Bouwer and Koster, 2016). For RQ1b, an alluvial plot was developed to visualize the results.

Main Analyses in View of RQ2

To investigate RQ2, a content analysis (on all texts, $n = 130$) and an in-depth analysis (on a subsample of texts, $n = 15$) were conducted. As to the content analysis, units of meaning were used to divide each text into multiple units of analysis. A unit of meaning can be a phrase, sentence or paragraph corresponding to one of the elements of an argumentative text (e.g., a rebuttal). The segmentation into units of analysis was executed by the first author. In total, 1,437 units of analysis were coded. Each unit of analysis is linked to one code, varying from 1 to 9. See **Table 5** for an overview of the codes assigned to each unit of analysis and an example of each code. **Table 5** is a representation of the code book that was developed. The code book provided detailed information concerning argumentative text characteristics and general text characteristics. To support the raters (the first author and a trained second rater), various examples and exceptions were also included in the code book.

In the content analysis, a second rater double-coded 24% ($n = 31$) of the collected, argumentative texts ($n = 130$). Within the 31 double-coded texts, 369 units of analysis were double-coded. Krippendorff's alpha was calculated to estimate the inter-rater reliability (Krippendorff and Hayes, 2007). The results indicate that the inter-rater reliability was high ($\alpha = 0.93$).

For the in-depth analysis, we selected a subsample of argumentative texts. By means of the preceding analyses of RQ1, several argumentative texts could be perceived as "mid-range," "weak" or "strong," independent of rating procedure used. For the purpose of this study, we define a mid-range text as a text in the middle 40–60% across rating procedures. A weak text is defined as a text in the lowest 20% of each rating procedure, and likewise, a strong text is a text that scores in the top 20% across rating procedures. All argumentative texts ($n = 130$) were ranked from highest to lowest for each rating procedure. We were then able to identify the top 20% and bottom 20% for each procedure. Next, it was examined which specific texts were always (regardless of rating procedure) in the top 20% (i.e., 7 texts) and bottom 20% (i.e., 7 texts). We applied the same process with the mid-range

TABLE 5 | Overview of the codes corresponding to each unit of meaning.

Code	Element	Example
Structure of argumentation		
1	Claim	*Zoos must be kept open.*
2	Claim data/argument	*Animals in zoos live longer and safer.*
3	Counterclaim	*Some people have the opinion that zoos should be closed.*
4	Counterclaim data	*As animals who are living in zoos are suffering from a lack of surface area.*
5	Rebuttal	*Living a longer and safer life is more important to me than having a lot of surface area.*
6	Rebuttal data	*By living longer and safer, almost extinct animal breeds have more opportunities to reproduce.*
General text characteristics		
7	Introduction	*The debate about whether zoos should close has been going on for some time. Several animal rights organizations have already taken action and protested. In this text, I will argue and defend my opinion on this conflict.*
8	Conclusion	*So from this I conclude that animals should actually be allowed to live in zoos.*
9	Irrelevant information	*I have already visited 5 zoos, situated in Antwerp, Brugelette, Mechelen, Vleteren and Ghent.*

texts. However, there was only one text that scored each time, across rating procedures, in the middle 40–60%. In this way, we arrived at the current selection of weak ($n = 7$), strong ($n = 7$) and mid-range ($n = 1$) texts.

This subsample of texts ($n = 15$) was subjected to an in-depth analysis in three different areas: (a) Structure of argumentation, (b) quality of content and (c) general textual characteristics. By means of this in-depth analysis, we try to uncover the characteristics of texts that have been scored as mid-range, weak and strong (see **Table 6**).

First, the structure of the mid-range, weak and strong argumentative texts was closely examined. Based on the content analysis, we checked which specific, argumentative elements were present in the mid-range, weak and strong argumentative texts. As mentioned in the see section "Data Analysis," an argumentative text ideally includes all argumentative elements. Earlier research also showed that often in weak argumentative texts, only a claim, and arguments supporting that claim, are provided implying that the author of the text is affected by tunnel vision and is ignoring the other point of view/counterclaim (Wolfe and Britt, 2008; Ferretti and Lewis, 2013). By means of the content analysis, we thus examined the argumentative structure of the fifteen texts in depth (e.g., how many of these texts consist only of a claim and claim data? If counterarguments are given, are they always refuted?). Second, the quality of the content was studied. In the theoretical background, we clarified that an argumentative text ideally has a strong persuasiveness, good factual accuracy and uses information originating from the source texts. All mid-range, weak and strong argumentative texts were analyzed on their quality of content by examining these three elements. See **Table 7** for more coding details. Third, to examine general textual information of this subsample of texts, the content analysis was used to determine whether the fifteen mid-range, weak and strong texts contain an introduction, conclusion, and/or irrelevant information. In addition, text length and writing mechanics (already taken into account in the preparatory analyses) were included in this in-depth analysis.

RESULTS

RQ1a: How Do the Three Rating Procedures Correlate?

Correlations between the three rating procedures are moderate to high, and are all significant at the 0.001 level (see **Table 8**). The correlations show that the different procedures are positively correlated, but are not fully aligned so they may focus on different text characteristics. Following Bouwer and Koster (2016), corrections on the correlations were conducted to deal with unreliability and to reflect the true correlations, as described in "Data Analysis" section.

TABLE 6 | In depth-analysis on mid-range, weak, and strong argumentative texts ($n = 15$).

Analyses on:	By means of:
Structure of argumentation	Content analysis
Quality of content	Analyses on persuasiveness, factual accuracy and use of information originating from source texts
General textual information	Content analysis and analyses on text length and writing mechanics

RQ1b: How Often Do We See Deviations Between Rating Procedures and How Strong Are These Deviations?

Knowing that correlations between the three rating procedures are moderate to high (and all significant at 0.001 level, it is interesting to inspect the descriptive statistics. In **Table 9** the descriptive statistics of the assigned scores using the three different rating procedures have been listed.

As the procedure of pairwise comparisons is our starting point from which connections are made to the other rating procedures (see section "The Present Study" of this study), a distinction was made between texts that were assigned a low score on pairwise comparisons (lowest 20%) but a high score (top 20%) on both of the other procedures (both analytic and holistic rating) and vice versa. First, not a single text was identified with a low score on pairwise comparisons, but a high score on the other rating procedures. Second, only one text was found rated as top 20% for pairwise comparison and bottom 20% for both holistic and analytic rating.

In an alluvial plot, the scores of the three rating procedures are compared to one another (see **Figure 1**). As can be observed, not all rating procedures arrive at the same ranking order. This indicates that each procedure has a specific focus. For instance, a text with a low holistic score does not necessarily have a low score on the other two rating procedures. Based on the inspection of the alluvial plot, there are some texts that are systematically ranked among the lowest or highest by all three of the rating procedures. However, there are also a large amount of texts that were evaluated rather differently by the three rating procedures indicating that in general the rankings fluctuate among the three rating procedures. This means that, though the correlations are in general positive and significant, the three rating procedures do not lead to exactly the same rankings.

RQ2: Which Elements Characterize Mid-Range, Weak, and Strong Texts?

Based on the in-depth analyses of the subsample, multiple elements characterizing argumentative texts were repeatedly identified among the mid-range, weak and strong texts. In the next section, the most common elements are described and analyzed.

Elements of a Mid-Range Argumentative Text

As we see many mid-range texts in each rating procedure, only one argumentative text was found which scored in the middle 40–60% each time, independent of the rating procedure being used. Mid-range argumentative texts do not have the highest scores but do not score particularly low either. In **Table 10**, the elements of the mid-range argumentative text are summarized. In general, the

TABLE 7 | Coding of quality of content.

Persuasiveness	Weak: The reasons the author puts forward are (a) not profound and (b) insufficient. Average: The reasons the author puts forward are (a) not profound or (b) insufficient. Strong: Deep reasoning from students. The author comes up with (a) profound and (b) sufficient reasons to support the claim.
Factual accuracy	Bad: Incorrect information was found at least twice in the text Average: Incorrect information was found once in the text Good: All information provided by the author was correct
Information originating from source texts	Never: No information in the author's text originated from the source texts Sometimes: Some information in the author's text originated from the source texts Always: All information in the author's text originated from the source texts

results reveal that the structure of this text was not great (i.e., no rebuttals were included, no rebuttal data were given), although the content and the general composition of the text was quite good. This explains why the text was, across all rating procedures, situated in the middle 20% (40–60%). As this is a single text, we will not go further into detail.

Elements Weak Argumentative Texts

Seven argumentative texts were found to score at the lowest 20% in the dataset according to all rating procedures. In **Table 11**, the elements of weak argumentative texts are summarized. If an element is observed in four or more out of the seven weak texts, we consider this a key element. The results reveal a weak argumentative text is characterized by: (a) The inclusion of only a claim and argument(s), (b) tunnel vision, (c) weak factual accuracy, (d) a lack of information from source texts, (e) weak persuasiveness, (f) the inclusion of irrelevant information, (g) short text length, and (h) weak writing mechanics.

Elements of Strong Argumentative Texts

Next to mid-range and weak texts, it was examined whether there were texts that are rated as the 20% strongest texts independent of rating procedure. Seven texts were found and, as can be seen in **Table 12**, several text features can be associated with a strong argumentative text. All these text features appeared in a minimum of four out of seven strong texts. More specifically, the results reveal that strong argumentative texts are characterized by (a) the use of a claim, arguments, a counterclaim, counterarguments, rebuttals, and rebuttal data, (b) all counterarguments are refuted by rebuttal(s), (c) the integration of information from source texts, (d) strong persuasiveness, (e) factual accuracy, (f) use of an introduction, (g) use of a conclusion, (h) high number of words, and (i) good writing mechanics.

TABLE 8 | Correlation coefficients between holistic rating, analytic rating and pairwise comparisons.

	Argumentative texts (n = 130)		
	Holistic	**Analytic**	**Pairwise comparisons**
Holistic	–	0.577**	0.483**
Analytic	0.913**	–	0.298**
Pairwise comparisons	0.818**	0.336**	–

***Correlation is significant at the 0.001 level. Above the diagonal are the original correlations, below the diagonal are the attenuated correlations.*

TABLE 9 | Descriptive statistics of the assigned scores using the rating procedures.

	HOLISTIC RATING (on a scale from 0 to 10)	**ANALYTIC RATING (on a scale from 0 to 100)**	**PAIRWISE COMPARISONS (Logit scores)**
M	5.77	50.19	0.07*
SD	1.44	19.36	1.135
Median	6	55	0.17
Minimum score	2	15	–3.46
Maximum score	9	85	2.47
Range	7	70	5.93

**In pairwise comparisons the mean of the logit scores is usually equal to zero. However, in this study, this score slightly deviates as the informative texts were left out of the ranking.*

DISCUSSION

The present study focused on providing insight into three different rating procedures by studying similarities (correlations) and deviations between scores assigned by each rating procedure. We argue that all three rating procedures are suitable for evaluating argumentative texts. However, when comparing the three procedures, we notice that in general, the rankings fluctuate among the three rating procedures. All three procedures can be seen as proxies for the quality of the argumentative texts, however, they have their own approach and focus. In addition, we found several elements of argumentative texts that seem to be associated with mid-range, weak or strong texts. In the discussion, these elements will be further explored. We aim to guide practitioners, researchers, and teachers in choosing a suitable rating procedure by verifying the purposes for which certain procedures work well. The discussion is structured according to the three research questions and, at the end, the findings are compiled and translated into practice. Limitations and suggestions for follow-up research are also discussed.

RQ1a: How Do the Three Rating Procedures Correlate?

Regarding the first research question (RQ1a), we found that the three rating procedures (i.e., absolute holistic rating, comparative

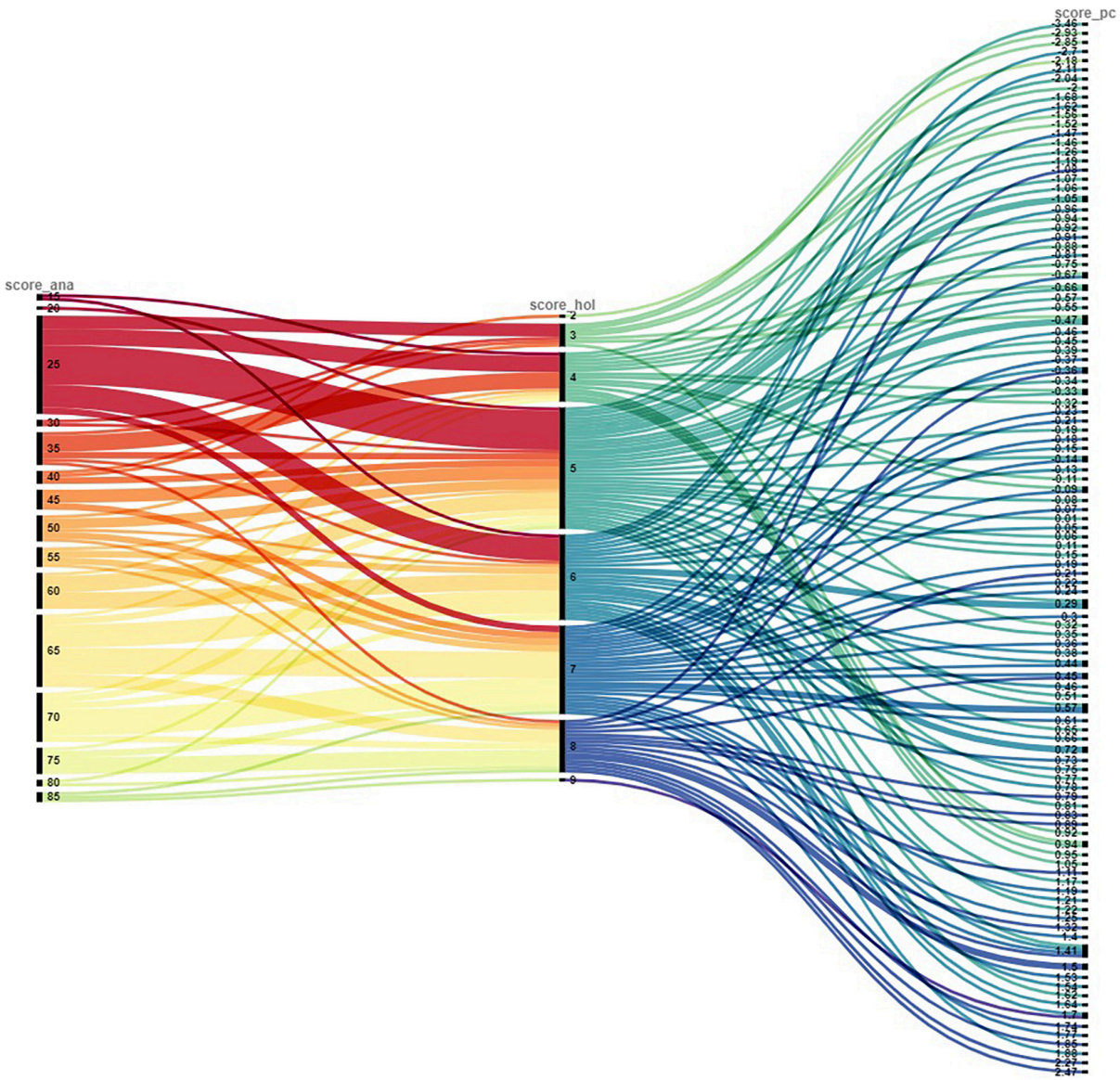

FIGURE 1 | Alluvial plot presenting the comparison of scores assigned through different rating procedures. Each colored line represents an argumentative text. The Figure shows a combination of the three procedures, with the analytic score on the left, the holistic score in the middle, and the pairwise comparisons score on the right. The scores per rating procedure are ordered from low (top of the alluvial plot) to high (bottom of the alluvial plot). Lines that are situated at the bottom of the figure represent a low score. Lines at the top of the figure represent a high score. The numbers on the vertical axes represent the scores attained by the rating procedure. Texts do not systematically score high or low on all three procedures, as becomes clear when looking at the figure.

holistic rating, and absolute analytic rating) correlate moderately to highly. Given that all procedures focus on assessing the quality of argumentative texts, this was in line with our expectations. However, the correlations are not fully aligned. Fully aligned correlations would indicate that rating procedures measure the underlying construct in exactly the same way (Messick, 1989). In this respect, the construct measured in this study is "argumentative writing skills."

This study revealed high attenuated correlations between absolute holistic and absolute analytic rating. When delving into the research literature, the findings on the correlations between both rating procedures are inconsistent. Studies by Freedman (1981) and Veal and Hudson (1983) show that holistic and analytic rating procedures correlate strongly. In contrast, studies of Hunter (1996) and Lee et al. (2009) indicate that holistic and analytic scores are not always and automatically strongly correlated. Keeping this contradiction in mind, in this study, we did not expect holistic and analytical rating to be this highly correlated because of the different focus of each procedure (i.e., holistic rating is based on the whole text whereas analytical rating focuses on specific argumentative text features). This high correlation could also be associated with the implementation of

TABLE 10 | Elements of the mid-range argumentative text.

	Elements mid-range argumentative texts
Quality of argumentation	Structure of the text: claim (1)—argumentation (3)—counterclaim (1)—counterargument (1) (no rebuttals nor rebuttal data)
Quality of content	Use of information from informative source texts Strong persuasiveness (but could be higher if there were counterarguments which were refuted by rebuttals) Good factual accuracy
General textual information	No introduction Conclusion No irrelevant information Average text length (250–400 words) Good writing mechanics

TABLE 11 | Elements of weak argumentative texts.

	Elements of weak argumentative texts	Number of texts (*n*)
Quality of argumentation	Only claim and arguments (no counterarguments nor rebuttals)	7
	Tunnel vision	7
Quality of content	Weak factual accuracy	4
	No use of information from source texts	5
	Weak persuasiveness	5
General textual information	Irrelevant information (= code 9)	5
	Short text length (<250 words)	4
	Bad writing mechanics	4

TABLE 12 | Elements of strong argumentative texts.

	Elements of strong argumentative texts	Number of texts (*n*)
Quality of argumentation	Use of (a) claim, (b) arguments, (c) counterclaim, (d) counterarguments, (e) rebuttal and (f) rebuttal data	4
	All counterarguments are refuted by rebuttal(s)	4
Quality of content	Use of information from source texts	7
	Strong persuasiveness	7
	Factual accuracy	6
General textual information	Introduction	4
	Conclusion	6
	Long text length (> 400 words)	7
	Good writing mechanics.	4

the rating procedures in the current study. More specifically, both the holistic and the analytical ratings were carried out by the same raters (see further in the section on limitations). The high attenuated correlation between absolute and comparative holistic rating was expected, as both are holistic procedures that look at the whole of the text. Alongside the high correlations between absolute holistic and absolute analytic rating and between absolute holistic and comparative holistic rating, the results in this study revealed rather low (but still significant at the 0.001 level) attenuated correlations between comparative holistic rating and absolute analytic rating. These results were expected given the different focus of the procedures (i.e., comparative holistic rating focuses on the whole text while analytic rating assesses different text features) and given the different underlying assessment strategy (i.e., analytic rating assesses texts in an absolute manner, while comparative holistic rating is based on comparing texts).

Based on the results of this study, the moderately to highly correlating rating procedures indicate the complexity of assessing argumentative texts. More specifically, argumentative writing is a complex interplay of various interrelated skills (such as reading skills, writing skills, and argumentation skills). Assessing such a complex and cognitively demanding activity requires assessment procedures that are able to grasp this complexity. The rating procedures central to this study each focused on assessing the quality of an argumentative text, resulting in relatively strong correlations, however they were far from perfectly aligned and, as the alluvial plot showed, the texts were ranked in different orders, which will be discussed in the next section.

RQ1b: How Often Do We See Deviations Between Rating Procedures and How Strong Are These Deviations?

Our findings showed that the rating procedures resulted in different ranking orders and that a text that is assigned a high score by one rating procedure, does not necessarily receive a high score by the other rating procedures. Given that the correlations are not fully aligned and as each rating procedure had its own focus of assessment (see RQ1a), this was expected. The deviations between the rating procedures were visualized in **Figure 1**. These findings reveal a certain level of agreement between the different procedures and indicate that despite different assigned scores, all procedures are suitable to assess the quality of an argumentative text.

We can conclude that the three rating procedures can be seen as proxies for the quality of argumentative texts, however, they have their own focus. Due to the nature of the analytic scoring process, the rating criteria in analytic rating are the most detailed. When all criteria are met, a high score is achieved, and although this is likely to result in high absolute and comparative holistic scores, this is not necessarily so. The opposite is true too: the best texts out of the comparative holistic approach might not necessarily have all elements required by the ASRAW. As all three procedures have their own focus, the scores will certainly not always be in line.

The conclusion that the texts are not exactly ranked in the same order by the three rating procedures should not necessarily be seen as a problem. It might be interesting to combine the different scores on one text, assigned by the different rating procedures, as feedback and input for the author. For example, as an author you can write a text that is assigned a low score by the ASRAW. An analytic rubric already offers opportunities for feedback: the author can clearly identify where points were lost (Bacha, 2001). But this same text could get a high score from comparative holistic rating (pairwise comparisons). The text then scores well in comparison to other texts written by peers. It can be interesting to look at texts written by peers: what can you

learn from these texts in terms of writing mechanics, transitions between paragraphs, text length, text structure, etc.? In light of feedback, it therefore seems interesting to combine the input of different assessment procedures.

RQ2: Which Elements Characterize Mid-Range, Weak and Strong Texts?

The results indicate that certain text features or elements seem to be associated with mid-range (see **Table 10**), weak (see **Table 11**) or strong argumentative texts (see **Table 12**). In this discussion, we will elaborate on the text elements that can be decisive in judging a text as a strong argumentative text. Several studies have investigated the quality of argumentation in students' writing. In this respect, previous studies have pointed out that many students do not include counterarguments and rebuttals in their argumentative texts (Wolfe and Britt, 2008; Ferretti and Lewis, 2013). Very often students only include a claim and claim data from their own point of view, resulting in a tunnel vision in which the opposite view is ignored (Nussbaum and Kardash, 2005). Ideally, all viewpoints should be recognized and supported but the opposite viewpoint should be less convincing than the chosen viewpoint, as Stapleton and Wu (2015) declare. In the present study, we confirmed the results of several previous studies (Figueredo and Varnhagen, 2005; Barkaoui, 2010; De La Paz and Felton, 2010; Rezaei and Lovorn, 2010; Stapleton and Wu, 2015; Cuevas et al., 2016; Jansen et al., 2021; Syed et al., 2021). These studies showed that the elements that seem to be associated with strong texts were: (a) Use of the (adapted) Toulmin elements, (b) refuting all counterarguments by rebuttal(s), (c) integrating information from source texts, (d) strong persuasiveness, (e) factual accuracy, (f) use of introduction and conclusion, (g) long text length, and (h) good writing mechanics. If the integration of the abovementioned elements is related to the overall text quality, we need to teach students how to integrate these text elements in their argumentative writing, as Wong et al. (2008) suggest. Furthermore, it is also important to be aware of these essential genre elements when assessing argumentative texts (regardless of which rating procedure is used). In this respect, we need to inform raters of these success criteria. In absolute analytic rating this can be done by using a rubric in which these elements are present; in absolute and comparative holistic rating we can inform the raters of the key elements of a good argumentative text by means of training.

For Which Purposes Do Certain Procedures Work Well?

All three rating procedures each have their own advantages, a different focus and different prerequisites. In this section, we aim to guide practitioners, researchers and teachers in choosing a suitable rating procedure for the writing assignment they have in mind. Given the variation in scores, it is important to consider when to use which rating procedure. In what follows, we will discuss the purposes for which certain procedures work well. We briefly sum up the situations in which each rating procedure can be used and we provide advantages and disadvantages.

Absolute Holistic Rating Procedure

When in need of a quick general score, absolute holistic rating is ideal as this is a very time-efficient procedure (Charney, 1984). Scores can be assigned by one rater, making this procedure particularly useful for teachers and practitioners. However, raters ideally have some experience in rating texts (Charney, 1984; Rezaei and Lovorn, 2010). Holistic rating was used in our research to assess argumentative texts, but this procedure can be used for other text genres as well. A disadvantage of absolute holistic rating is that validity and reliability cannot be ensured (Wesdorp, 1981; Charney, 1984). The present study confirmed these previous studies as the absolute holistic rating procedure had a rather low reliability. However, this might be a problem for empirical researchers, but teachers and practitioners may value the quickness and naturalness of this procedure. In addition, we could address the low reliability by giving raters more guidance and training (Charney, 1984), e.g., in using the whole scoring range. In this respect, other absolute holistic assessments can also be implemented, e.g., a holistic rubric instead of general impression marking may help to obtain more reliable scores (Penny et al., 2000).

Comparative Holistic Rating Procedure (Pairwise Comparisons)

Pairwise comparisons use multiple raters to develop a rank order from lowest to highest text quality. Consequently, the need for multiple raters makes it difficult to implement this rating procedure in daily practice. However, as Bouwer et al. (2018) claim, assessing competences through pairwise comparisons is an easier task than using an analytic rubric which precisely pays attention to multiple text features. As high validity and reliability can be achieved, this procedure is very interesting for empirical researchers. Neither absolute holistic nor analytic rating automatically guarantee reliability, as we discussed above. A reliable rating procedure will, if applied again, obtain similar results in a following measurement (Charney, 1984). In our findings, we achieved an SSR of 0.83 for our pairwise comparisons. Researchers or practitioners that choose to use this procedure should pay attention to the provided instruction. It is possible that raters pay equal attention to quality of content, quality of argumentation, and general textual information. However, this cannot be fully assured: You cannot be sure in advance whether assessors will pay equal attention to these elements. Raters can always be influenced by their own thoughts on what defines a good text. Special attention should also be paid to text length, as our research demonstrated that longer texts were often rated as more qualitative texts. This may be due to the fact that text length is an easy, holistic criterion to use as this is the first visual indicator raters see when they are presented with two texts to compare (Lee et al., 2009).

Analytic Rating Procedure

In contrast to comparative holistic rating, analytic rating is workable for one person, making this a procedure that can be useful for teachers and practitioners. In addition, the analytic rating procedure can achieve high reliability (in our research: ICC = 0.98), but this is not automatically the case. Earlier

research on reliability of analytic rating is still inconsistent. Therefore, Harsch and Martin (2013) suggest combining holistic and analytic rating procedures to achieve more reliable and valid results. In contradiction to the holistic rating procedure, training a rater is less important as raters only have to decide the category in which a certain text feature can be put, without further justification. However, we do not claim that analytic rating is always easy; deciding the level in which a certain feature belongs can be a difficult choice to make when there is doubt. In addition, raters need a clear view on the argumentative elements when analytically rating argumentative texts. Identifying claims or counterclaims is not self-explanatory. The absolute analytic rating procedure, and more specifically the ASRAW, can only be used when the main focus is on the structure of the argumentation. In this research, the ASRAW-rubric was used to assign scores to argumentative texts. Of course, other instruments can also be used to analytically score argumentative texts. The ASRAW mainly focuses on the quality of argumentation. If the structure of the argumentation is not good, the final score is automatically low. However, the ASRAW does pay attention to the quality of content, but only after taking a closer look at the structure of the argumentation. Writing mechanics and text length are not included in the ASRAW-rubric and therefore seemed to have less impact here compared to pairwise comparisons.

Limitations and Suggestions for Further Research

In what follows, limitations regarding the implementation of the rating procedures are addressed. In addition, suggestions for further research are proposed.

A first limitation focuses on the low reliability (ICC = 0.48) of the holistic rating procedure. We used general impression marking as the absolute holistic rating procedure, but it could have been interesting to use other absolute holistic assessment methods (e.g., holistic rubrics) as they provide additional resources to raters to assign a score to a text. As Weigle (2002) points out: "the scoring procedures are critical because the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Therefore, other assessment methods within holistic rating are also a possibility. We cannot guarantee that using a different holistic rating procedure would have had a positive effect on the ICC, but research by Penny et al. (2000) indicates that higher rater agreement could be achieved by means of using a holistic procedure containing additional support for raters.

The second limitation relates to the implementation of the rating procedures. Two out of three rating procedures (i.e., absolute holistic and analytic rating) were conducted by the first author and a second trained rater. Both raters rated the same texts holistically as well as analytically, which could partly explain correlations between the two procedures. The first author and the trained rater first implemented the holistic rating procedure, followed by the analytic rating procedure. This could have influenced the assessments, however, there was 1 week in between the ratings and the analytic ratings were implemented in a different order than the holistic ratings to avoid interdependency. For future research studies, we recommend that raters do not rate the same text holistically as well as analytically. Regarding validity and reliability, Harsch and Martin (2013) prefer rating a text both holistically and analytically. In providing feedback to students, this could be very useful. However, research by Hunter (1996) and Lee et al. (2009) showed that holistic and analytic scores are not always strongly correlated. More research is needed into the implications of merging information from both holistic rating and analytical rating. In our opinion, using both the holistic and the analytic rating procedure can indeed be a suitable way to assess texts, as Harsch and Martin (2013) suggest, but in practice it may not always be time-efficient and manageable to apply multiple rating procedures.

A third limitation focuses on the training of the raters. For holistic and analytic rating, both raters were intensively trained, unlike the 133 university students that conducted pairwise comparisons. The university students received a very short briefing, but no extensive training like the two raters that rated analytically and holistically was provided. The university students were no experts in assessing argumentative writing skills. Given the comparative nature of the writing assessment in the pairwise comparisons, we opted not to interfere so the university students could rely on their overall knowledge of argumentative writing, based on the provided broad criteria. Therefore we gave only few instructions to the university students. We notice some discrepancies in the educational literature on rating procedures. On the one hand, more general assessment studies of Sadler (1989) and Pollitt (2012) suggest that experienced raters can assess texts more easily, because of their experience. On the other hand, recent research on pairwise comparisons suggests that comparing texts is a relatively simple task and that rater experience is therefore not necessary (van Daal et al., 2016; Coertjens et al., 2017). On this view, pairwise comparisons may also work without an extended training. From this, we can conclude that training raters could have an influence on differences between the rating procedures, but for pairwise comparisons, little training should be sufficient in order to get reliable results. In addition, differences between raters cannot always be solved by training them in advance (Coertjens et al., 2017).

A fourth limitation relates to the 27 texts that were omitted from the study as they were informative instead of argumentative. These texts did not take a position (e.g., pro or contra), did not have the goal to persuade, nor were any arguments integrated. While this was a deliberate decision for research purposes, it is, however, not feasible in practice, as teachers cannot omit texts from evaluation. In a classroom context, all texts (whether argumentative or not) should be evaluated by the teacher.

CONCLUSION

The research field on writing assessment generally distinguishes between holistic and analytic rating procedures. However, another distinction has been recently identified: Absolute and

comparative rating procedures (Bouwer and Koster, 2016; Coertjens et al., 2017). To date, there is little research that focuses on both distinctions. Therefore, this study is one of the first studies comparing absolute holistic rating, with comparative holistic rating (pairwise comparisons) and absolute analytic rating. In this study, we especially focus on the more innovative approach of pairwise comparisons, as this procedure is compared to more established methods of absolute holistic and analytic rating. In this study, it was indicated that the three rating procedures correlate moderately to highly, but each have different qualities, advantages and prerequisites. However, all three procedures are suitable for practitioners to use when assessing argumentative texts. In addition, we focused in detail on the deviance between the three rating procedures and the characteristics of mid-range, weak, and strong argumentative texts.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Faculty of Psychology and Educational Sciences of Ghent University (Specific Ethical Protocol for Scientific Research). Written informed consent for participation was not provided by the participants' legal guardians/next of kin because: There was an active written informed consent from the participants and a passive written informed consent from participants' parents.

AUTHOR CONTRIBUTIONS

YL, FD, BD, and HV designed the study. YL was in charge of the data collection procedure. YL analyzed the data, with the help of FD, BD, and HV. All authors wrote and reviewed the manuscript and approved its final version.

FUNDING

This research was supported by a grant from the Research Foundation Flanders (FWO) (Grant No. G010719N).

REFERENCES

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System* 29, 371–383. doi: 10.1016/S0346-251X(01)00025-2

Barkaoui, K. (2010). Explaining ESL essay holistic scores: a multilevel modeling approach. *Lang. Test.* 27, 515–535. doi: 10.1177/0265532210368717

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assess. Edu. Princ. Policy Pract.* 18, 279–293. doi: 10.1080/0969594X.2010.526585

Bouwer, R., and Koster, M. (2016). *Bringing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students*. Utrecht: Universiteit van Utrecht.

Bouwer, R., Goossens, M., Mortier, A. V., Lesterhuis, M., and De Maeyer, S. (2018). *Een comparatieve aanpak voor peer assessment: leren door te vergelijken. Toetsrevolutie: Naar Een Feedbackcultuur in Het Hoger Onderwijs*. Culemborg: Uitgeverij Phronese. 92–106.

Bramley, T., Bell, J. F., and Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Edu. Res. Persp.* 25, 1–24.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issu. Appl. Ling.* 11:35. doi: 10.5070/l4112005035

Clark, D., and Sampson, V. (2007). Personally-seeded discussions to scaffold online argumentation. *Int. J. Educ. Sci.* 3, 351–361. doi: 10.1080/09500690600560944

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Res. Teach. Eng.* 18, 65–81.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., and De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of *via* paarsgewijze vergelijking: Een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische Studien* 94, 283–303.

Cuevas, I., Mateos, M., Martín, E., Luna, M., and Martín, A. (2016). Collaborative writing of an argumentative synthesis from multiple sources: the role of writing beliefs and strategies to deal with controversy. *J. Writ. Res.* 8, 205–226. doi: 10.17239/jowr-2016.08.02.02

De La Paz, S., and Felton, M. K. (2010). Reading and writing from multiple source documents in history: effects of strategy instruction with low to average high school writers. *Contemp. Edu. Psychol.* 35, 174–192. doi: 10.1016/j.cedpsych.2010.03.001

Ferretti, R. P., and Lewis, W. E. (2013). "Best practices in teaching argumentative writing," in *Best practices in writing instruction* 2nd ed, eds S. Graham, C. A. MacArthur, and J. Fitzgerald (New York, NY: Guilford Press), 113–140.

Figueredo, L., and Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Read. Psychol.* 26, 441–458. doi: 10.1080/02702710500400495

Follman, J. C., Anderson, J. A., and Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading English themes. *Res. Teach. Eng.* 1, 190–200. doi: 10.1111/j.1365-2214.2011.01355.x

Freedman, S. (1981). Influences on evaluators of expository essays: beyond the text. *Res. Teach. Eng.* 15, 245–255.

Gill, T., and Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assess. Edu. Princip. Policy Pract.* 20, 308–324. doi: 10.1080/0969594X.2013.779229

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *J. Res. Dev. Edu.* 27, 73–82.

Graham, S., and Perin, D. (2007). What we know, what we still need to know: teaching adolescents to write. *Sci. Stud. Read.* 11, 313–335. doi: 10.1080/10888430701530664

Granado-Peinado, M., Mateos, M., Martín, E., and Cuevas, I. (2019). Teaching to write collaborative argumentative syntheses in higher education. *Read. Writ.* 32, 2037–2058. doi: 10.1007/s11145-019-09939-6

Harsch, C., and Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assess. Edu. Princip. Policy Pract.* 20, 281–307. doi: 10.1080/0969594X.2012.742422

Hunter, D. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *Canad. J. Prog. Eval.* 11, 61–85.

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *Colleg. Comp. Commun.* 41:201. doi: 10.2307/358160

Huot, B. A. (1993). "The influence of holistic scoring procedures on reading and rating student essays," in *Validating holistic scoring for writing assessment: theoretical and empirical foundations*, eds M. M. Williamson and B. A. Huot (New York, NY: Hampton Press).

Jansen, T., Vögelin, C., Machts, N., Keller, S., and Möller, J. (2021). Don't just judge the spelling! the influence of spelling on assessing second-language student essays. *Front. Learn. Res.* 9:44–65. doi: 10.14786/flr.v9i1.541

Jarvis, S., Grant, L., Bikowski, D., and Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *J. Second Lang. Writ.* 12, 377–403. doi: 10.1016/j.jslw.2003.09.001

Kibler, A. K., and Hardigree, C. (2017). Using Evidence in l2 argumentative writing: a longitudinal case study across high school and university. *Lang. Learn.* 67, 75–109. doi: 10.1111/lang.12198

Krippendorff, K., and Hayes, A. F. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Measur.* 1, 77–89. doi: 10.1080/19312450709336664

Lee, Y. W., Gentile, C., and Kantor, R. (2009). Toward automated multi-trait scoring of essays: investigating links among holistic, analytic, and text feature scores. *Appl. Ling.* 31, 391–417. doi: 10.1093/applin/amp040

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison-Welsley.

Marttunen, M., Laurinen, L., Litosseliti, L., and Lund, K. (2005). Argumentation skills as prerequisites for collaborative learning among finnish, french, and english secondary school students. *Edu. Res. Eval.* 11, 365–384. doi: 10.1080/13803610500110588

McMahon, S., and Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assess. Edu. Princip. Policy Pract.* 22, 368–389. doi: 10.1080/0969594x.2014.978839

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Edu. Res.* 18, 5–11. doi: 10.3102/0013189X018002005

Myers, M. (1980). *A Procedure for Writing Assessment and Holistic Scoring. In College Composition and Communication*. Urbana, IL: National Council of Teachers of English and Educational Resources Infor- mation Center

NCES (2012). *The Nation's Report Card: Writing* 2011. Washington, DC: NCES.

Nussbaum, E. M., and Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *J. Edu. Psychol.* 97, 157–169. doi: 10.1037/0022-0663.97.2.157

Nussbaum, E. M., and Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *J. Exp. Edu.* 76, 59–92. doi: 10.3200/JEXE.76.1.59-92

Penny, J., Johnson, R. L., and Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: an empirical study of a holistic rubric. *Assess. Writ.* 7, 143–164. doi: 10.1016/S1075-2935(00)00012-X

Pollitt, A. (2012). Comparative judgement for assessment. *Int. J. Technol. Design Edu.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

Qin, J., and Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System* 38, 444–456. doi: 10.1016/j.system.2010.06.012

Rezaei, A. R., and Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assess. Writ.* 15, 18–39. doi: 10.1016/j.asw.2010.01.003

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instruct. Sci.* 18, 119–144. doi: 10.1007/BF00117714

Sasaki, M., and Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Lang. Test* 16, 457–478. doi: 10.1177/026553229901600403

Simon, S. (2008). Using Toulmin's Argument Pattern in the evaluation of argumentation in school science. *Int. J. Res. Method Edu.* 31, 277–289. doi: 10.1080/17437270802417176

Song, Y., and Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. *Read. Writ.* 26, 67–90. doi: 10.1007/s11145-012-9381-8

Stapleton, P., and Wu, Y. (2015). Assessing the quality of arguments in students' persuasive writing: a case study analyzing the relationship between surface structure and substance. *J. Eng. Acad. Purp.* 17, 12–23. doi: 10.1016/j.jeap.2014.11.006

Syed, S., Al-Khatib, K., Alshomary, M., Wachsmuth, H., and Potthast, M. (2021). Generating informative conclusions for argumentative texts. *arXiv* doi: 10.48550/arXiv.2106.01064

Thorndike, E. (1920). A constant error in psychological ratings. *J. Appl. Psychol.* 4, 25–29. doi: 10.1037/h0071663

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/0033-295X.101.2.266

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Traga Philippakos, Z. A., and MacArthur, C. A. (2019). Integrating collaborative reasoning and strategy instruction to improve second graders' opinion writing. *Read. Writ. Quart.* 2019, 1–17. doi: 10.1080/10573569.2019.1650315

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Edu. Princip. Policy Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

van Rijt, J. H. M., van den Broek, B., and De Maeyer, S. (2021). Syntactic predictors for text quality in Dutch upper-secondary school students' L1 argumentative writing. *Read. Writ.* 34, 449–465. doi: 10.1007/s11145-020-10079-5

Varghese, S. A., and Abraham, S. A. (1998). Undergraduates arguing a case. *J. Second Lang. Writ.* 7, 287–306. doi: 10.1016/S1060-3743(98)90018-2

Veal, R. L., and Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Res. Teach. Eng.* 17, 290–296. doi: 10.3390/v13081651

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Measur.* 42, 428–445. doi: 10.1177/0146621617748321

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

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs*. Den Haag: Staatsuitgeverij.

Wolfe, C. R., and Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Think. Reason.* 14, 1–27. doi: 10.1080/13546780701527674

Wolfe, E. W., Song, T., and Jiao, H. (2016). Features of difficult-to-score essays. *Assess. Writ.* 27, 1–10. doi: 10.1016/j.asw.2015.06.002

Wong, B. Y. L., Hoskyn, M., Jai, D., Ellis, P., and Watson, K. (2008). The comparative efficacy of two approaches to teaching sixth graders opinion essay writing. *Contemp. Edu. Psychol.* 33, 757–784. doi: 10.1016/j.cedpsych.2007.12.004

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., and Baek, S. Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC Med. Edu.* 18:1–6. doi: 10.1186/s12909-018-1228-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.