frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/feduc.2022.823895

Validity of Comparative Judgment Scores: How Assessors Evaluate Aspects of Text Quality When Comparing Argumentative Texts

Marije Lesterhuis[1], Renske Bouwer[2], Tine van Daal[1], Vincent Donche[1] and Sven De Maeyer[1]*

[1] Edubron, Department of Training and Educational Sciences, University of Antwerp, Antwerp, Belgium, [2] Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, Netherlands

The advantage of comparative judgment is that it is particularly suited to assess multidimensional and complex constructs as text quality. This is because assessors are asked to compare texts holistically and to make a quality judgment for each text in a pairwise comparison based upon on the most salient and critical differences. Also, the resulted rank order is based on the judgment of all assessors, representing the shared consensus. In order to be able to select the right number of assessors, the question is to what extent the conceptualization of assessors prevails in the aspects they base their judgment on, or whether comparative judgment minimizes the differences between assessors. In other words, can we detect types of assessors who tend to consider certain aspects of text quality more often than others? A total of 64 assessors compared argumentative texts, after which they provided decision statements on what aspects of text quality had informed their judgment. These decision statements were coded on six overarching themes of text quality: argumentation, organization, language use, language conventions, source use, references, and layout. Using a multilevel-latent class analysis, four different types of assessors could be distinguished: narrowly focused, broadly focused, source-focused, and language-focused. However, the analysis also showed that all assessor types mainly focused on argumentation and organization, and that assessor types only partly explained whether the aspect of text quality was mentioned in a decision statement. We conclude that comparative judgment is a strong method for comparing complex constructs like text quality. First, because the rank order combines different views on text quality, but foremost because the method of comparative judgment minimizes differences between assessors.

Keywords: comparative judgment, validity, writing assessment, assessor cognition, latent class analysis

OPEN ACCESS

Edited by:
Anders Jönsson,
Kristianstad University, Sweden

Reviewed by:
Stephen Humphry,
University of Western Australia, Australia
Wei Shin Leong,
Ministry of Education, Singapore

****Correspondence:***
Marije Lesterhuis
Marije.Lesterhuis@uantwerpen.be

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

Received: *28 November 2021*
Accepted: *21 March 2022*
Published: *13 May 2022*

Citation:
Lesterhuis M, Bouwer R, van Daal T, Donche V and De Maeyer S (2022) Validity of Comparative Judgment Scores: How Assessors Evaluate Aspects of Text Quality When Comparing Argumentative Texts. Front. Educ. 7:823895. doi: 10.3389/feduc.2022.823895

INTRODUCTION

Comparative judgment is particularly suited for the judgment of complex skills, competencies and performances, such as writing or mathematical problem solving (Pollitt and Crisp, 2004; Heldsinger and Humphry, 2010; Pollitt, 2012; Jones et al., 2015). A characteristic of the assessment of complex skills is that the quality of students' work cannot be considered as either right or wrong, but on a continuum of quality. The quality is determined by

multiple aspects that are highly intertwined (Sadler, 2009). For text quality this is for example the content, structure, style and grammar. To judge the quality, human judgment is key. However, it is important that scores reflect the complexity of the skill under assessment. Therefore, the way assessors come to a judgment within comparative judgment plays a major role in its validity argument (Bejar, 2012).

Within comparative judgment, assessors base their judgment on holistic interpretations of the quality of students' work. That means that the assessor makes a single comparison of which text is better considering the communicative effectiveness. In other words, to what extent is a text reaching its communicative goal? These holistic interpretations take into account the complexity of what quality comprises (Lesterhuis et al., 2019). Also, when comparing, assessors can rely on their expertise and their own conceptualization of quality. This supports the validity of the assessment scores, because final scores are based on all judgments made by all assessors and thus represent the shared consensus on what quality comprises (Jones and Inglis, 2015). Previous empirical studies have indeed shown that assessors focus on different aspects when they make comparisons. Hence, involving a group of assessors in comparative judgment enhances construct representation (Pollitt and Whitehouse, 2012; Whitehouse, 2012; van Daal et al., 2019).

Yet, there is still little insight in the number and type of assessors that should be involved for valid comparative judgments and the role assessors play to achieve full construct representation (Messick, 1989; Lesterhuis et al., 2019). Therefore, we need to know the extent to which assessors differ from each other regarding the probability that certain aspects of quality are assessed. Up to now, studies only focused upon differences between assessors (van Daal et al., 2019) but didn't look for different profiles or types of assessors. Therefore, this study investigates whether assessors focus on different or similar aspects of students' argumentative texts when making comparative judgments, and to what extent different types of assessors can be distinguished based on their judgments. These insights help future assessment coordinators with the selection of assessors in order to achieve text scores that can be validity interpreted as representing the quality of the texts.

BACKGROUND

To understand the role that assessors play in a valid interpretation of text scores, this section discusses how assessors make comparative judgments, the types of assessors that can be distinguished with respect to the aspects of text quality they value and previous studies within the field of comparative judgment that looked into the aspects assessors base their decisions on.

Assessors' Judgment Process

The assessment of text quality requires the assessors to translate the text' quality into a judgment on that quality. In case of comparative judgment this is the decision which text is of higher quality. Therefore, an assessor reads the two texts, interprets the texts considering the different aspects and formulates a judgment on the quality of the whole. Assessors' cognition or mental scheme determines the way the assessors go through the text and how they conceptualize the texts' quality. Especially the latter is important for a valid interpretation of text scores because it affects what kind of aspects assessor do and do not value when assessing text quality. Consequently, the results of pairwise comparisons are based on assessors' conceptualization of text quality.

However, assessors can differ in how they conceptualize text quality (Huot, 1990; Vaughan, 1991). Various studies that look into how assessors judge single texts have investigated possible causes of this difference. For example, Cummings and others found that second language raters pay more attention to language than to rhetoric ideas, in contrast to first language raters. Wolfe (1997) found that proficient raters focus more on general features. Wang et al. (2017) found that inexperienced assessors differ in how they consider textual borrowing, the development of ideas, and the consistency of the focus. Consequently, all these studies show that assessors differ in how they translate texts into scores. Therefore, key in the validity argument is understanding how within a scoring method as comparative judgments, the selection of assessors affects construct representation.

Differences Between Assessors and Assessor Types

Studies on the assessment of text quality based on holistic and analytic scoring show that the aspects assessors consider is not fully random, but that assessors tend to belong to a certain type. For instance, Diederich et al. (1961) asked assessors to score 300 texts holistically on a nine-point scale without any instructions or criteria; meanwhile, the assessors had to provide the texts with written comments. The assessors differed to a large extent regarding the quality level to which they assigned the texts. Based on these differences, the researchers classified the assessors into five groups. Additionally, the researchers analyzed the assessors' comments and examined whether the comments differed between the groups. All assessors had focused on the clarity of expression, coherence, and logic (reasoning). However, the groups differed in the importance they attached to the relevance, clarity, quantity, development, and soundness of ideas (idea-focused); on organization and spelling (form-focused); on style, interest and sincerity (creativity-focused), on the errors in texts (mechanics-focused); on the choice and arrangement of words (effectiveness-focused). Based on this study, we can expect that assessors differ in what aspects of text they value while comparing two texts and consequently also the way they score text quality.

To look into the effect of conceptualization of text quality on analytic scoring with criteria, Eckes (2008) analyzed the importance that 64 experienced assessors attributed to nine quality criteria. He identified six groups of assessors. Four groups were more-or-less like the groups identified by Diederich et al. (1961); the groups focused on syntax, correctness, structure, and fluency. However, Eckes also found two groups that could be typified according to the aspects they considered less important

compared to the other assessor types: not fluency-focused and not argumentation-focused. In a follow-up study, Eckes (2012) showed that the groups were related to how these assessors scored texts on a rating scale. He found that belonging to a certain type of assessors relates to how severe an assessor rates a criterion.

Using a similar approach, Schaefer (2016) found three groups of assessors when analyzing how 40 relatively untrained English teachers scored 40 English essays using criteria. He distinguished the assessors that focused on rhetorical features, linguistic features and mechanics. Schaefer (2016) could, however, not substantiate the link between the aspects the assessors said they valued and the aspects they really valued when scoring texts. Nevertheless, these studies showed that assessors differ in how they conceptualize text quality and that this affects how outcomes, in this case scores, using holistic or criteria-based scoring methods.

Differences Between Assessors in Comparative Judgment

A main advantage of comparative judgment is that assessors only have to provide relative decisions. Consequently, differences in severity (i.e., one assessor systematically giving lower scores) do not affect the reliability of the results anymore. Most studies indeed show that the resulting rank order of the comparative judgments of multiple assessors is highly reliable. Using 10 till 14 comparisons per text (or other types of student work) generates already acceptable reliability estimates (Separation Scale Reliability = SSR) of 0.70 (Verhavert et al., 2019). A high SSR reflects a high stability in the way the texts are ranked (Verhavert et al., 2018). This is a prerequisite for valid scores. However, valid scores also require that these quality scores fully reflect the complex construct of text quality.

To what extent do assessors take the full construct of text quality in account when making comparative judgment? Some empirical studies have already investigated the aspects that assessors consider when choosing texts by looking into how assessors justify their decision. In the study of van Daal et al. (2019), the explanations of 11 assessors to justify their decisions when comparing academic papers were analyzed. Analysis of these explanations—or decision statements—revealed that the group of assessors considered all relevant aspects of text quality and did not consider irrelevant aspects as the basis of their comparative judgments. Also, all assessors focused predominantly on the structure of the text and source use. However, there was still considerable variance between assessors. They varied in the aspects they also considered and in the number of aspects they mentioned in their decision statements. For example, some assessors focused on the discussion section, while others did not, and some mentioned language errors, while others did not. In the study of Lesterhuis et al. (2019), 27 teachers compared argumentative texts, referring to a wide range of aspects of text quality when justifying their decisions, varying from aspects of the argumentation to whether a title was present. However, whether assessors differ systematically on the aspects they discriminate on when comparing texts, or whether this has been caused by the different text pairs has not been established.

Humphry and Heldsinger (2019) show that the texts that are in a certain pair inform what assessors consider. They asked assessors to tick which aspects of 10 criteria informed their judgment after making a comparison. They found that when assessors compare lower quality texts, assessors more often base their decision on sentence structure and spelling and grammar and when comparing texts of higher quality, they referred to audience orientation and setting and character. This raises the question whether there are trends among assessors in the aspects they consider when comparing texts, independent of the pair of texts. In other words, can different types of assessors be detected when looking at the aspects assessors refer to when justifying their decision?

RESEARCH AIMS

Previous research focusing on other scoring methods has shown that assessors develop different conceptualizations of text quality which affect how they judge the quality of texts. It is yet unknown whether assessors take different aspects into account when making a comparison decision. This is relevant because in the method of comparative judgments, assessors can play a major role in the aspects that are considered because they are not forced to assess text quality on predefined quality criteria (as is the case in analytic judgments), but instead they can rely on their own expertise when comparing texts in a holistic manner (e.g., Pollitt, 2012; Jones et al., 2015). Previous studies already suggest that assessors make comparative judgments based on a wide range of relevant aspects of text quality, showing that the shared consensus of the resulting rank-order reflects the complexity of the construct of text quality (van Daal et al., 2017; Lesterhuis et al., 2019). This does not fully reveal whether assessors are comparable or whether different types of assessors exist, and hence, multiple assessors are needed for a valid assessment. Therefore, the central question in this study is whether different types of assessors can be distinguished that tend to base their comparative judgments on certain aspects. And when types of assessors can be distinguished, how can we typify these classes? These insights are important to understand the role of assessors in the validity argument and how the selection of assessors affects a valid interpretation of text scores.

MATERIALS AND METHODS

Participants

To search for trends among assessors, we chose a varied selection of assessors. A total of 64 assessors participated in this study. They had an average age of 37.23 years (SD = 14.22), 20 were men, 44 were women, and all were native Dutch speakers. Of the 64 assessors, 32.8% were student teachers, with no experience teaching or evaluating students' work; 42.2% were teachers (years of experience M = 19.96, SD = 13); 14.1% were teacher trainers (years of experience M = 13.11, SD = 7.67); and 9.4% worked as examiners (years of experience M = 23, SD = 9.17) working for an

organization that certifies students who are following an irregular educational track.

The Assessment

The assessors evaluated the quality of three argumentative writing tasks completed by 135 students at the end of secondary education in their first language (Dutch). The students had to write an argumentative essay about the following topics: "Having children," "Organ donation," and "Stress at school." These tasks were previously used in the research of van Weijen (2009) but were adjusted slightly to the Flemish context. For each task, the students received six short sources, which they had to use to support their arguments. We included three tasks, so the findings do not depend on one specific task.

The tasks were in line with the competence "argumentative writing" as formulated in the final attainment goals of the Flemish Department of Education.[1] These goals were familiar to all assessors and students, and described what students need to be able to at the end of secondary education. The students had 25 min for each task. The 135 texts with the topics "Having children" and "Organ donation" were used. However, due to practical issues, only 35 randomly selected texts with the topic "Stress at school" were included in this study.

Procedure

Assessors came together on the campus, two times for 2 h. Before starting the assessment, the assessors received an explanation about the method of comparative judgment. Also, we gave a short introduction of the students' tasks and the competence of argumentative writing. The Digital Platform for the Assessment of Competences tool (D-PAC) supported the assessments that used comparative judgment. Within this tool, three assessments were created, each including texts of only one topic. For the topic "Having children," 1,224 pairs were generated; for the topic "Organ donation," 901 pairs were generated; and for the topic "Stress at school," 474 pairs were generated. In total, 2,599 comparative judgments were made. These pairs were randomly assigned to the assessors, who started with the assessment of "Having children," followed by "Organ donation" and "Stress at school." For each pair, the assessors decided which of the two texts was of higher quality in light of the competence of argumentative writing.

Next, they responded after each comparison to the query "Can you briefly explain your judgment?" Based on these decision statements, information was gathered on the aspects of text quality that informed the decisions of the assessors (Whitehouse, 2012; Bartholomew et al., 2018; van Daal et al., 2019). Each assessor made at least 10 comparisons, with a maximum of 56 comparisons (M = 40.60, SD = 16.16). The variation in the number of comparisons was due to assessors only attended one judgment session and/or because of differences in judgment speed. The assessors provided a decision statement for 98,1% of the made comparisons.

[1] www.onderwijsdoelen.be

Pre-analysis

Using user-defined functions in R, we applied the Bradley-Terry-Luce model to the data (Bradley and Terry, 1952; Luce, 1959), in order to estimate logit scores for each text. These logit scores express the log of the odds, which can be transformed to the probability that a particular text will be selected as the better text when compared to a text of average quality. These scores can be interpreted as a text quality score. The reliability of these scores was calculated by taking the variation in quality and the standard error of each text's quality score. This reliability is expressed in the scale separation reliability (*SSR*). The texts with the topic "Having children" had an *SSR* of 0.81, "Organ donation" was 0.73 and "Stress at school" 0.89. These high reliabilities show that the comparative judgments across the assessors were consistent (Verhavert et al., 2019).

Analyses

All decision statements were coded according to seven aspects of text quality, argumentation, organization, language use, formal language conventions, source use and references. A total of 10% of the assessment "having children" was double coded and showed a sufficient level of reliability of K = 0.65 (Stemler, 2004). **Table 1** shows the percentage each element was mentioned according to all the assessments and assessors.

In order to detect whether types of assessors can be distinguished, a data file was created in which it was indicated for each comparison whether the assessor had mentioned an aspect of text quality (1) or not (0). A multilevel latent class (MLCA) analysis was performed on this dataset, as comparisons were nested in assessors. A latent class analysis investigates if there are trends in the answers given by assessors, by examining the probability of an aspect being mentioned by an assessor. Assessors with the same probability of mentioning an aspect are grouped in a class (Vermunt and Magidson, 2003). By describing this class, a type of assessor is created.

In order to determine how many classes of assessor types can be distinguished, several models are estimated. Each model contains one class more than the previous model. To select the best fitting model, we looked first into the Bayes Information Criterion (*BIC*) and the total Bivariate Residuals (*TBVR*). For both the *BIC* and *TBVR*, we were interested in the relative reduction, which indicates the importance of adding another class to the class solution regarding model fit (van den Bergh et al., 2017).

Second, we used the classification error and entropy (E) to investigate the different class solutions. The classification error refers to the certainty that each assessor can be assigned to one of the distinguished classes. The classification error increases when several assessors show a high probability of belonging to more than one class. The entropy is a single number summary of the certainty with which assessors can be assigned to a class. This depends, on the one hand, on the overlap of classes with regard to their probability patterns and, on the other hand, on how well assessors can be assigned to a single class according to their modal posterior probabilities. The closer the entropy

TABLE 1 | Coding scheme with example statements and percentages elements were mentioned.

Aspect of text quality	Example statement	% mentioned in decision statement (*N* = 2,599)
Argumentation	The arguments used in the left text are better supported (comparison 164, a teacher with 3 years of experience)	57.3
Organization	I think the organization and form of the structure are better (comparison 227, a teacher trainer with 6 years of experience)	56.0
Language use	Beautiful and surprising use of language is important when you aim to convince someone (comparison 359, a teacher with 7 years of experience)	23.4
Formal language conventions	However, this text has grammar and construction mistakes (comparison 1,977, a teacher with 30 years of experience)	19.2
References	The references to sources are done well (comparison 2,309, a student with no experience)	19.2
Source use	Better integration of the sources (comparison 1,713, a student with no experience)	18.2
Layout	. . . The aspects of the text quality that are seen easily, as layout, length and the presence of a title	17.0

is to 1, the more certain assessors can be assigned to class (Collins and Lanza, 2009).

When the best fitted model of classes is selected, the differences between classes will be described by a horizontal and a vertical analysis. This description results in different types of assessors. Using the Wald statistic, we examine whether a class differs significantly from other classes with regard to aspects that are mentioned (horizontal analysis). In addition, for each class we will look at the probability that particular text quality aspects are mentioned in the decision statements (vertical analysis).

Because experience and occupational background can be related to how assessors conceptualize text quality, we checked whether these assessors' characteristics were related to the class composition. Experience has been operationalized as the assessors' number of years of relevant experience of teaching and/or writing assessment. To check the relationship with the group composition, the Welch test was executed, because the number of years did not meet the assumption of equal variances between groups. The occupational background was operationalized as an assessor being a student teacher, teacher, teacher trainer or examiner. To check the relationship with group composition, a chi-square test was performed.

RESULTS

Exploring the Number of Assessor Classes

Table 2 shows that the *BIC* and the *TBVR* kept deteriorating by adding a class to the class solution. However, the relative increase of the model fit stopped after the four-class solution.

The four-class solution also appeared to be good when investigating the certainty that assessors could be assigned to a class. According to the classification error and the entropy presented in **Table 2**, the four-class solution resulted in a better assignment of assessors to classes than the two-, three-, or five-class solution, as the classification error was 0.02 and the entropy 0.96. To illustrate, when applying a four-class solution, 61 assessors can be assigned to a class with a probability exceeding 90%. For the remaining three assessors, the highest probability to belong to a class is 77, 71, and 59%. Based on these results, we argue that assessors can be divided into four homogeneous sub-classes concerning the probability that they refer to an aspect of text quality.

TABLE 2 | Model parameters and classification of assessors to classes.

Model	*BIC* (*LL*)	Δ *BIC* (*LL*)	*TBVR*	Classification error	*E*
One class	19,912.42	46.64	–	1	
Two classes	19,447.51	−464.91	36.34	0.02	0.93
Three classes	19,236.24	−211.27	30.91	0.02	0.95
Four classes	19,118.89	−117.36	27.62	0.02	0.96
Five classes	19,022.71	−96.17	24.74	0.03	0.95
Six classes	18,937.78	−84.93	22.26	0.01	0.97
Seven classes	18,863.06	−74.73	19.90	0.02	0.97
Eight classes	18,816.70	−46.36	18.08	0.03	0.96

Describing the Differences Between Assessor Classes

This section describes the class solution in greater depth. It begins with a general description of the four-class solution and is followed by a description of each of the four classes.

The Class Solution

The best class solution divided the 64 assessors into four assessor classes. The classes differed in size, however, each class consisted of a substantial number of assessors. The first class consisted of 35.06% ($n = 22$) of the assessors, the second class 32.52% ($n = 21$), the third class 18.74% ($n = 12$), and the fourth class 13.68% ($n = 9$).

The four classes differed significantly in each aspect of text quality that they mentioned, as shown by the Wald tests ($W \geq 52.95$, $p < 0.01$) and in the average number of aspects they mentioned in a decision statement [Welch's $F(3,1062.12) = 243.17$, $p < 0.01$]. The R^2 in **Table 3** shows that the extent that this class' solution explained whether an aspect of text quality was mentioned, varied between 0.02 for argumentation and 0.11 for the layout. In other words, although differences between the classes are significant, the class solution does not fully explain whether a particular aspect of text quality was mentioned in the decision statements.

TABLE 3 | Explanatory power of the four-class solution.

	R^2
Argumentation	0.02
Organization	0.08
Language use	0.06
Source use	0.09
Language conventions	0.03
References	0.07
Layout	0.11

Four Types

To describe the classes in depth, **Table 4** reflects the differences between the classes when using the Wald statistic with a paired comparison approach. **Figure 1** visualizes the probability an aspect of text quality was mentioned by each class.

Class 1 contained the largest number of assessors (n = 22) and can be indicated as language-focused. This class referred most often to the organization of texts, and subsequently to the argumentation. However, typical for this class is that it additionally referred regularly to language use and language conventions, with 36 and 26%, respectively. For language use, this probability is significantly higher than the other classes. For language conventions, the probability is higher than the assessors in class 2 and class 3. Moreover, this class referred to 2.43 (SD = 1.13) aspects of text quality in a decision statement, on average.

Class 2 (n = 21) can be called narrowly focused. Only argumentation and organization were deemed to be relevant to these assessors. However, with 48% for argumentation and 41% for organization, this class did not even refer to these aspects regularly, compared to the other classes. The narrow focus is also reflected in the number of aspects mentioned, on average, in each decision statement (M = 1.38, SD = 0.94). The Games-Howell *post hoc* test showed that this was significantly less than the other classes ($p < 0.01$).

Class 3 (n = 12) can be indicated as source-focused. Besides the argumentation and organization of texts, this class found source use and references to be the most important aspects for choosing a text. The 34% probability for source use is significantly higher than the other classes, and the 28% probability for references is significantly more than classes 1 and 2. This class reflected on 2.16 (SD = 1.14) aspects per decision statement on average.

Class 4 (n = 9) can be typified as broadly focused. In addition to argumentation, this class was more likely to refer to all aspects than at least two of the other classes. Moreover, each aspect was mentioned with more than a 25% probability. That broad focus was also reflected in 3.04 (SD = 0.94) aspects that this class averagely mentioned in a decision statement. According to the Games-Howell *post hoc* test, this is significantly more than

TABLE 4 | The probability that assessors within a class refer to an aspect of text quality.

Class 1 language-focused		Class 2 narrowly focused	Class 3 source-focused	Class 4 broadly focused	Average probability
Argumentation	0.63	0.48*	0.66	0.55*	0.57
C2: $W = 36.42$, $p < 0.01$ *C3*: $W = 1.27$, $p = 0.2e6$ *C4*: $W = 6.43$, $p = 0.01$		*C3*: $W = 36.64$, $p < 0.01$ *C4*: $W = 4.07$, $p = 0.04$		*C4*: $W = 10.04$, $p < 0.01$	
Organization	0.71	0.41	0.46	0.71	0.56
C2: $W = 148.11$, $p < 0.01$ *C3*: $W = 79.74$, $p < 0.01$ *C4*: $W = 0.00$, $p = 0.96$		*C3*: $W = 2.61$, $p = 0.11$ *C4*: $W = 78.39$, $p < 0.01$		*C4*: $W = 48.10$, $p < 0.01$	
Language use	0.36*	0.13	0.15	0.28*	0.23
C2: $W = 113.64$, $p < 0.01$ *C3*: $W = 63.30$, $p < 0.01$ *C4*: $W = 7.25$, $p < 0.01$		*C3*: $W = 0.75$, $p = 0.39$ *C4*: $W = 33.53$, $p < 0.01$		*C4*: $W = 19.20$, $p < 0.01$	
Source use	0.16*	0.06*	0.34	0.33	0.18
C2: $W = 40.24$, $p < 0.01$ *C3*: $W = 54.03$, $p < 0.01$ *C4*: $W = 42.55$, $p < 0.01$		*C3*: $W = 144.21$, $p < 0.01$ *C4*: $W = 123.18$, $p < 0.01$		*C4*: $W = 0.03$, $p = 0.86$	
Language conventions	0.26	0.11*	0.15*	0.27	0.19
C2: $W = 56.60$, $p < 0.01$ *C3*: $W = 20.27$, $p < 0.01$ *C4*: $W = 0.08$, $p = 0.78$		*C3*: $W = 4.30$, $p = 0.03$ *C4*: $W = 40.15$, $p < 0.01$		*C4*: $W = 16.09$, $p < 0.01$	
References	0.16*	0.10*	0.28*	0.40*	0.19
C2: $W = 14.49$, $p < 0.01$ *C3*: $W = 25.83$, $p < 0.01$ *C4*: $W = 71.60$, $p < 0.01$		*C3*: $W = 67.41$, $p < 0.01$ *C4*: $W = 125.16$, $p < 0.01$		*C4*: $W = 12.75$, $p < 0.01$	
Layout	0.15	0.09*	0.12	0.49*	0.17
C2: $W = 11.82$, $p < 0.01$ *C3*: $W = 2.15$, $p = 0.14$ *C4*: $W = 131.85$, $p < 0.01$		*C3*: $W = 144.21$, $p < 0.01$ *C4*: $W = 123.18$, $p < 0.01$		*C4*: $W = 117.43$, $p < 0.01$	

**Significantly different from all other classes with $p < 0.05$.*

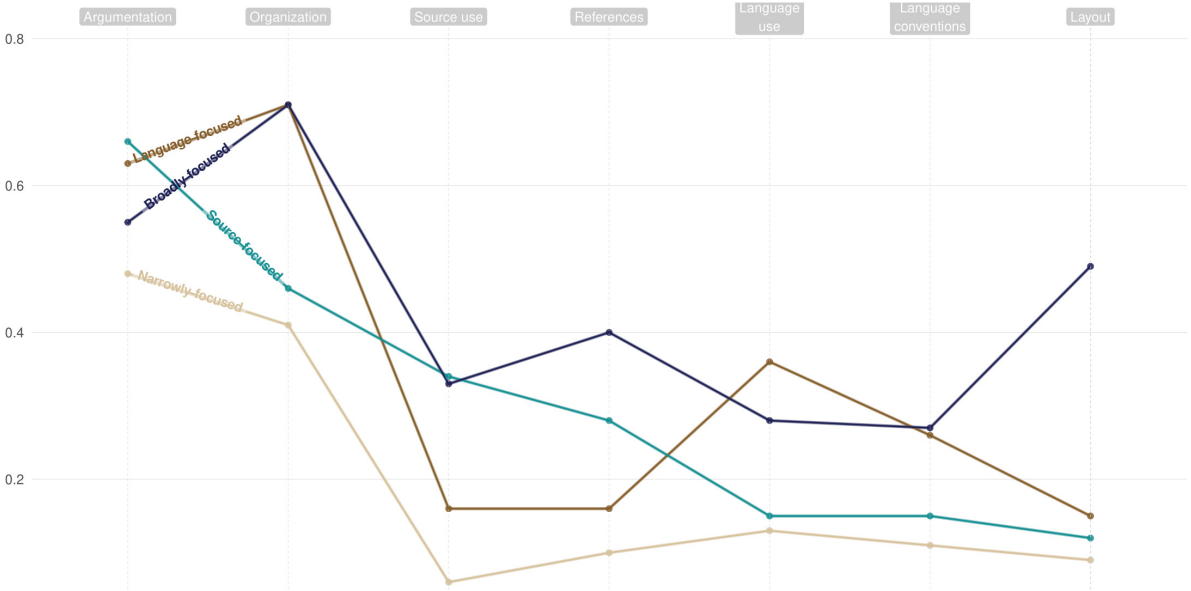

FIGURE 1 | The probability that a class referred to an aspect of text quality in a decision statement.

the average number of aspects mentioned by the other classes ($p < 0.01$).

Controlling for Experience

As the assessors differed in relevant years of experience and occupational background, we investigated whether these assessor characteristics related to the distinguished types. A Welch test showed that years of relevant experience had no significant effect on the composition of classes [$F(3, 25.84) = 0.91$, $p = 0.45$]. Next, the chi-square test showed there is no statistically significant relationship between occupational background and the classes [χ^2 (12, $N = 63$) = 9.31, $p = 0.68$].

DISCUSSION

Comparative judgment is especially suited to assess complex skills. As assessors are assumed to vary in the aspects upon which they focus, combining their judgments should foster construct representation (Pollitt and Whitehouse, 2012; Whitehouse, 2012; van Daal et al., 2019). However, it is unclear whether differences between assessors occur systematically. Therefore, this study examined to what extent types of assessors can be discerned. A type of assessor refers to a group of assessors that systematically considers an aspect of text quality (or not) when discerning between two texts. To investigate whether different types of assessors could be distinguished we analyzed 2,599 decision statements that 64 assessors gave to explain their comparative judgments on the quality of argumentative texts of students in the fifth grade of secondary education. These decision statements were coded on argumentation, organization, language use, source use, language conventions, references, and layout. Next, we applied a MLCA to investigate whether classes of assessors with a similar argumentation pattern could be detected.

Based on the MLCA, four classes of assessors could be distinguished. All assessor classes referred to organization and argumentation when making comparative judgments but differed with regards to other aspects of text quality. Class 1 was mainly language-focused. These assessors were more likely to mention language use and conventions to justify their comparative judgments than the other classes of assessors. Class 2 was narrowly focused, which means that assessors in this class hardly referred to other aspects in a decision statement than argumentation and organization. Class 3 was source-focused, assessors within this class were more likely to focus on source use and references. Class 4 was broadly focused, these assessors considered a great number of aspects of text quality when comparing texts.

The types of assessors are in line with research using absolute scoring procedures, where content and organization were mostly considered when assessing text quality (Vaughan, 1991; Huot, 1993; Sakyi, 2003; Wolfe, 2006). The language-focused class was related to the classes distinguished by Diederich et al. (1961) and Eckes (2008). Moreover, the source-focused class underpins Weigle and Montee's (2012) result that only some assessors consider the use of sources when assessing text quality. However, we did not find the same types of assessors as the other studies. This raises the question whether the method (comparative judgment) or the type of writing task impacted the determined types of assessors. For instance, in contrast to this study, the tasks used by Diederich et al. (1961); Eckes (2008), and Schaefer (2016) did not require the use of sources. This could explain the fact that a source-focused class was only found in our study, but not in other studies on rater types. Studies on how assessors adjust their focus according to the task they assess will improve our understanding of the stability of the types of assessors across tasks.

It is important to note that assessors were instructed to assess the full construct of text quality (argumentation, organization,

language use, source use, language conventions, references, and layout). This study showed that for the validity of text scores, multiple assessors should be involved in a comparative judgment assessment. This increases the probability that the multidimensionality of text quality is represented in the text scores. To illustrate, the narrowly focused class rarely chose a text due to the quality of the source use, whereas the broadly- and source-focused class did. The latter group, however, rarely chose a text because of the language aspects. Combining the judgments of different types of assessors into scores leads to more informed text scores, representing the full complexity of text quality. Reading all these aspects of text quality in the decision statement underpins the argument that the final rank-orders represent the full construct of text quality in these assessments. In other assessments it might be of less importance that all of these aspects are considered. This depends on the aim of the assessment. For example, in some cases an assessment does not aim to assess students on the extent they apply language conventions. Most important for the validity argument is that students and assessors are aware of the assessment aim, so the aspects that are assessed by assessors are aligned with the student assignment.

Interestingly, the explanatory power of the types on the aspects that were assessed was rather limited. For example, all classes referred mostly to argumentation and organization and the four classes explained only for 2% whether argumentation was referred to in a decision statement (11% for layout). Also, none of the classes referred to one of the aspects of text quality in each and every decision statement, for example, the broadly focused class referred to organization in 71% of the decision statements. That means that for none of the classes one of the aspects of text quality was always the reason to choose for one text over the other. This conclusion seems to underpin the hypothesis that by offering assessors a comparison text, they rely less on their internalized ideal text but that their judgment is affected by the specific texts they are comparing. It makes us aware that more research is needed to establish what factors are at play. Pollitt and Murray (1996), Bartholomew et al. (2018), and Humphry and Heldsinger (2019) suggested that the quality of student works is related to the aspects upon which assessors focus. Comparing lower quality performance, lower order aspects as grammar and sentence structure seem to be more salient to assessors, whereas when comparing higher quality performance, the stylistic devices and audience are. Future research should consider both the assessor and characteristics of the text pair when looking into what aspects inform the comparison.

More research is also needed to better understand the implications for the resulting rank order of texts. Do assessors who belong to the same type make the same decisions on which texts are better? And does this differ from assessors belonging to another type? Studies on holistic and analytic scoring methods showed the relationship between what aspects were considered and resulting text scores (e.g., Eckes, 2012). But it is unknown whether this link can also be established within the context of comparative judgment. Unfortunately, the current data collection does not provide sufficient data to calculate text scores per assessor class.

The decision statements were shown to be a rich data source, enabling the detection of systematic differences between assessors. They were gathered during the assessment and did not interfere with the judgment process to a great extent. However, they only provided insight into the aspects that assessors revealed they based their judgments on, the just-noticeable-difference. They did not reveal information on the judgment process. To triangulate and extend the conclusions of this study, other data sources are required. Specifically, using think-aloud protocols while assessors make the comparisons would help us to understand what aspects assessors focus on when reading the texts, and how this relates to the aspects they subsequently base their decision on (Cumming et al., 2002; Barkaoui, 2011). This would enable us to gain insight into whether the narrowly focused and broadly focused classes also take other processing actions to reach a judgment. For example, Vaughan (1991) and Sakyi (2000) found that some assessors take only one or two aspects into account before deciding using an absolute holistic scoring procedure. Within the context of comparative judgment, this way of making decisions seemed to be typical for the whole narrowly focused class. Additionally, the broadly focused class, on average, referred to more aspects of text quality in a decision statement. This result suggests that these assessors apply a more analytical approach when comparing texts. Research into whether these differences in decision statements really reflect different processing strategies is needed to design comparative judgment in such a manner that it would optimally support assessors to make valid judgments.

CONCLUSION

We have concluded that different types of assessors can be distinguished based on differences in the aspects that the assessors were more likely to base their judgment on when comparing texts. These types have, however, only a small explanatory power regarding what aspects are assessed and all assessor' types had their main focus on organization and argumentation.

Nevertheless, the fact that we could detect assessor types implies that texts are ideally compared by multiple assessors—with different perspectives on text quality. Moreover, comparative judgment has been shown to be a promising way to integrate the judgments of multiple assessors into valid and reliable scores of text quality.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are made available by the authors, without undue reservation through OSF (Lesterhuis et al., 2022).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ML and SD were responsible for the data collection. ML, RB, VD, and SD were responsible for the research design and conceptualization of the research questions. ML was responsible for the analyses. ML, RB, and TD were responsible for drafting the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Flanders Innovation and Entrepreneurship and the Research Foundation under (Grant No. 130043).

REFERENCES

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Lang. Test.* 28, 51–75. doi: 10.1177/0265532210376379

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., and Reeve, E. M. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educ. Assess.* 23, 85–101. doi: 10.1080/10627197.2018.1444986

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Meas. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika* 39, 324–345. doi: 10.1093/biomet/39.3-4.324

Collins, L. M., and Lanza, S. T. (2009). *Latent Class And Latent Transition Analysis: With Applications In The Social, Behavioral, And Health Sciences*, Vol. 718. Hoboken, NJ: John Wiley & Sons.

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Mod. Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

Diederich, P. B., French, J. W., and Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Res. Bull. Ser.* 1961:98.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Lang. Test.* 25, 155–185. doi: 10.1177/0265532207086780

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Lang. Assess. Q.* 9, 270–292. doi: 10.1080/15434303.2011.649381

Heldsinger, S., and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/bf03216919

Humphry, S., and Heldsinger, S. (2019). Raters' perceptions of assessment criteria relevance. *Assess. Writ.* 41, 1–13. doi: 10.1016/j.asw.2019.04.002

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *Coll. Composit. Commun.* 41, 201–213. doi: 10.1097/00001888-199404000-00017

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 (Creskhill, NJ: Hampton Press, Inc), 206236.

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math.* 89, 337–355. doi: 10.1007/s10649-015-9607-1

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., and De Maeyer, S. (2022). *Validity of Comparative Judgment Scores [dataset]*. OSF. doi: 10.17605/OSF.IO/8X692

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., and De Maeyer, S. (2019). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Lit.* 18:1. doi: 10.17239/L1ESLL-2018.18.01.02

Luce, R. D. (1959). On the possible psychophysical laws. *Psychol. Rev.* 66, 81–95.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189x018002005

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594x.2012.665354

Pollitt, A., and Crisp, V. (2004). "Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?," in *Proceedings of the British Educational Research Association Annual Conference, UMIST, Manchester, September 2004* (Cambridge: UCLES).

Pollitt, A., and Murray, N. L. (1996). "What raters really pay attention to," in *Performance Testing, Cognition and Assessment*, eds M. Milanovic and N. Saville (Cambridge: University of Cambridge), 7491.

Pollitt, A., and Whitehouse, C. (2012). *Using Adaptive Comparative Judgement To Obtain A Highly Reliable Rank Order In Summative Assessment*. Manchester: AQA Centre for Education Research and Policy.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assess. Eval. High. Educ.* 34, 159–179. doi: 10.1080/02602930801956059

Sakyi, A. A. (2000). "Validation of holistic scoring for ESL writing assessment: How raters evaluate," in *Fairness And Validation In Language Assessment: Selected Papers From The 19th Language Testing Research Colloquium, Orlando, Florida* ed. A. J. Kunnan (Cambridge: Cambridge University Press), 129.

Sakyi, A. A. (2003). *Validation of Holistic Scoring for ESL Writing Assessment: How Raters Evaluate Compositions*. Ph.D. thesis. Toronto, ON: University of Toronto.

Schaefer, E. (2016). "Identifying rater types among native english-speaking raters of english essays written by japanese university students," in *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* eds V. Aryadoust, and J. Fox (Cambridge: Cambridge Scholars), 184.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract. Assess. Res. Eval.* 9:4.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Princ. Policy Pract.* 26, 59–74. doi: 10.1080/0969594x.2016.1253542

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M. T., Donche, V., and De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: the moderating role of decision accuracy. *Front. Educ.* 2:44. doi: 10.3389/feduc.2017.00044

van den Bergh, M., Schmittmann, V. D., and Vermunt, J. K. (2017). Building latent class trees, with an application to a study of social capital. *Methodology* 13, 13–22. doi: 10.1027/1614-2241/a000128

van Weijen, D. (2009). *Writing processes, Text Quality, And Task Effects: Empirical Studies In First And Second Language Writing.* Dissertation, Netherlands Graduate School of Linguistics, Amsterdam.

Vaughan, C. (1991). "Holistic assessment: What goes on in the rater's mind," in *Assessing Second Language Writing In Academic Contexts* ed. L. Hamp-Lyons (Norwood, NJ: Ablex), 111–125.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

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: 10.1177/0146621217748321

Vermunt, J. K., and Magidson, J. (2003). Latent class models for classification. *Comput. Stat. Data Anal.* 41, 531–537.

Wang, J., Engelhard, G., Raczynski, K., Song, T., and Wolfe, E. W. (2017). Evaluating rater accuracy and perception forintegrated writing assessments using a mixed-methods approach. *Assess. Writ.* 33, 36–47. doi: 10.1016/j.asw.2017.03.003

Weigle, S. C., and Montee, M. (2012). "Raters perceptions of textual borrowing in integrated writings tasks," in *Measuring Writing: Recent insights into Theory, Methodology and Practice*, eds E. VanSteendam, M. Tillema, G. C. W. Rijlaarsdam, and H. van den Bergh (Leiden: Koninklijke BrillNV), 117145.

Whitehouse, C. (2012). *Testing The Validity Of Judgements About Geography Essays Using The Adaptive Comparative Judgement Method*. Manchester: AQA Centre for Education Research and Policy.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometricscoring system. *Assess. Writ.* 4, 83–106. doi: 10.1016/S1075-2935(97)80006-2

Wolfe, E. W. (2006). Uncovering raters cognitive processing and focus using think-aloud protocols. *J. Writ. Assess.* 2, 37–56.

Conflict of Interest: ML and SD were co-founders of Comproved. However, this company was only founded after the research was conducted, analyses executed and first draft written (as a chapter in a dissertation).

The remaining 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.