WHEN TEACHERS COMPARE ARGUMENTATIVE TEXTS

Decisions informed by multiple complex aspects of text quality

MARIJE LESTERHUIS*, TINE VAN DAAL*, ROOS VAN GASSE*, LIESJE COERTJENS**, VINCENT DONCHE* & SVEN DE MAEYER*

** University of Antwerp ** Psychological Sciences Research Institute, Université Catholique de Louvain*

Abstract
Insight into aspects that guide teachers' decisions when assessing student text quality is crucial to an understanding of the validity of text scores. Such research has been lacking in the context of comparative methods which are, however, increasingly being used for text assessment purposes. This study reports on the aspects of argumentative texts that guide Flemish teachers' decisions when using the comparative judgement method. In using this method, teachers indicate which text of a pair is of higher quality. 27 teachers explained 23 comparative judgements in a decision statement, when comparing randomly selected texts written by 135 students in their fifth year of general secondary education. This resulted in 596 statements referring to 2054 segments of aspects of text quality. Firstly, an inductive analysis revealed that teachers consider a wide range of aspects with regard to text quality when making comparison decisions. Secondly, the deductive aggregation of these aspects showed that most decisions are informed by the organisation and argumentation of the texts. Lastly, almost all statements reported complex aspects of text quality, whereas half of the decision statements also showed a reflection on the rule-applying aspects of text quality. We conclude that comparative judgement encourages teachers to make decisions on complex and multiple aspects of text quality. Further research should elaborate on whether the aspects that informed teachers' decisions are related to the text they choose, and whether teachers differ in the aspects to which they refer.

Keywords: text assessment, comparative judgement, teachers' decision, argumentative writing, validity

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V. & De Maeyer, S. (2018). When teachers compare argumentative texts. Decisions informed by multiple complex aspects of text quality. L1-Educational Studies in Language and Literature, 18, pp. 1-22. https://doi.org/10.17239/L1ESLL-2018.18.01.02

Corresponding author: Marije Lesterhuis, University of Antwerp, Faculty of Social Sciences, Department of Training and Educational Sciences, Gratiekapelstraat 10, B-2000, Antwerp, Belgium, email: Marije.Lesterhuis@uantwerpen.be

1. INTRODUCTION

The assessment of texts is an important part of writing education. It informs teachers and students on the extent to which learning objectives have been attained. However, the assessment of texts is not easy, due to text quality being a multidimensional and complex construct (Messick, 1989). It contains lower order aspects such as the application of language conventions, through to more complex and higher order aspects such as structuring content (Cumming, Kantor, & Powers, 2002). In order to obtain valid scores when assessing the quality of texts, the scoring method should take this complexity into account (Messick, 1994; Moss, 1994). Because in most scoring methods it is up to the assessor to judge the text quality, investigating which aspects inform assessors' decisions is important in terms of validity claims for any assessment of writing (Cumming et al., 2002; Kane, 2006; Messick, 1994).

In assessment practices, different methods have been developed to cope with the complexity of assessing text quality; from absolute holistic or analytic methods, to, more recently introduced, comparative holistic methods. Absolute holistic scoring requires assessors to decide on the overall quality of a single text. This method has the advantage of being efficient (Schipolowski & Boehme, 2016; Weigle, 2002). In addition, proponents of holistic scoring appraise the explicit role of assessors as readers, in which their personal and overall reaction to texts is valued (Huot, 1990; White, 1984). At the same time, this causes problems in terms of obtaining reliable scores, as assessors value different aspects of text quality differently (Cai, 2015; Weigle, 2002). This has led to a more analytic approach to scoring, focusing on procedures that enhance the transparency of the aspects of text quality that assessors should consider. In an analytical approach, the construct of text quality is unraveled into multiple separate quality aspects. When assessors evaluate a text, it is assumed that they pay attention to these aspects individually (Weigle, 2002). However, it is difficult to identify all criteria beforehand and in a concrete manner. In addition, various studies have shown that often assessors start with holistic evaluations, which inform their scoring with regard to individual criteria (Cumming et al., 2002; Sakyi, 2003; Vaughan, 1991; Wolfe, 1997). Moreover, when a text does not fit these criteria perfectly, assessors often adjust their scores (Lumley, 2002). Consequently, the discussion has been raised as to whether or not unraveling text quality into separate aspects might lead to a loss in validity (Pollitt & Crisp, 2004; Sadler, 1989, 2009b).

As a reaction to these problems, more holistic procedures are becoming increasingly favored, combining these with the power of comparing rather than scoring texts absolutely. These methods support assessors' holistic scoring by providing them with texts to compare. These texts can be exemplar texts representing certain levels of quality (Bouwer, Béguin, Sanders, & van den Bergh, 2015), or other texts that also need to be scored. The latter leads to relative measures of text quality, as is the case with comparative judgement. For this method, the texts are presented to the assessors in pairs. For each pair, assessors indicate which text is of higher quality. To gather enough information about the quality of each text, texts are randomly compared with multiple other texts. The quality of each text is then calculated by analyzing all the comparative judgements made by all assessors (more information on the method can be found in Lesterhuis, Verhavert, Coertjens, Donche & De Maeyer, 2016, and Pollitt, 2012a, 2012b). Given that comparing texts is an easier and more reliable task when it comes to assessing complex issues such as text quality (Gill & Bramley, 2013), comparative methods have been increasingly appraised (Bramley, 2007; Heldsinger & Humphry, 2010; Pollitt, 2012a; Steedle & Ferrara, 2016). However, little is known about which aspects of text quality assessors consider when comparing texts, which hampers a valid interpretation of the assessment outcomes.

Nevertheless, the aspects that assessors focus on, or value most, can have an impact on the decisions they make (Bloxham, den-Outer, Hudson, & Price, 2016; Diederich, French, & Carlton, 1961; Eckes, 2012; Pollitt & Murray, 1996). In addition, different scoring methods encourage assessors to look at other aspects of texts. For example, in the case of analytic scoring, the rating scales attracts particular attention during scoring, whereas in holistic scoring the text is more central (Barkaoui, 2010). Holistic scores are less task-specific, and therefore more generalizable than analytic scores (Bouwer et al., 2015; Schoonen, 2005; van den Bergh, De Maeyer, van Weijen, & Tillema, 2012). To date, not much is known about which aspects guide assessors' decisions when using comparative methods. This is problematic for a complex construct such as text quality, because the comparative approach might encourage assessors to focus on superficial aspects (e.g., layout), or view the text rather narrowly (only taking one aspect into account) (Bramley, 2007).

Therefore, this study aims to achieve a better understanding of the aspects of text quality that inform assessors' comparison decisions, and how the aspects that inform teachers' decisions can be elucidated. This is important in order to understand how comparative methods enable teachers to assess complex constructs such as text quality. We will focus on the assessment of argumentative texts by teachers in the context of Flanders (the Dutch-speaking part of Belgium), which will be discussed in the next section. This section will also describe the construct of argumentative writing. Next, we will explain how the method of comparing texts is used to elicit the aspects that teachers value when making relative decisions.

2. ARGUMENTATIVE WRITING AND THE CONTEXT OF FLANDERS

The ability to write is important for students' educational careers (Preiss, Castillo, Grigorenko, & Manzi, 2013), especially since students are often assessed in terms of their ability to write argumentative texts (Brown, 2010; Hirvela, 2017; McColly, 1970; Wingate, 2012). Important in argumentative writing is the development of an argument, which requires students to take a position (often in relation to sources) and provide warrants and backings for this position. Students need to be able to present the components of an argument in a logical and coherent manner. In addition, the language component, concerning style and conventions, is relevant in determining the quality of the text (Bachman & Palmer, 2010; Ryder, Vander Lei, & Roen, 1999). However, the assessment of argumentative writing in general, and the use of sources in an argumentative text more specifically, is not straightforward (Gebril & Plakans, 2014; Wang, Engelhard, Raczynski, Song, & Wolfe, 2017; Weigle & Montee, 2012). Various studies have elaborated on what a good argumentative text should look like, and how students learn to write such texts. However, language researchers and specialists can only agree that there is no single best description of text quality in general (Chapelle, 2008). We also see that teachers have difficulties in conceptualizing which aspects determine the quality of an argumentative text, which can hamper them in teaching students to write these texts (Wingate, 2012).

Within Flanders, teachers develop their own assessment procedures, thereby determining the prerequisites for students to enter the next educational level. This high degree of autonomy is, among others, reflected in the non-existence of national exams or tests. Consequently, teachers do not share standards on a regular basis. However, the Flemish government has formulated final attainment goals for general and subject-related domains to guide teachers in the level they should aim for with their students. These goals cover the minimum requirements for different grades, and are described in terms of knowledge, skills and attitudes. Writing is part of both the generic domain and the subject-related domain of language (Dutch). The generic goals regarding writing prescribe that students need to be able to use their knowledge and information, in order to develop a logic text in which attention has being paid to content and functional relationships, to apply content and form conventions of language, pay attention to layout and use citations correctly. With regard to argumentative writing, the subject-related goals prescribe that students should be able to write integrated and argumentative texts for an (un)known audience on a judgmental level (http://eindtermen.vlaanderen.be).

These goals describe the components of text quality on a general level. It is unclear whether and which of these aspects inform teachers' comparison decisions. However, it is important to understand the validity of comparative methods within the context of Flanders. Concerns have been raised as to whether providing a national framework is sufficient for teachers to start using it (Skar & Jølle, 2017; Wyatt-Smith, Klenowski, & Gunn, 2010). Therefore, insight into how teachers use these goals when assessing texts is required. In addition, it must be asked whether or not teachers in today's secondary schools are too much focused on rule-applying aspects such as spelling and grammar (Van Grinsven, Mondrian, & Westerik, 2007; Wingate, 2012). By unravelling the aspects teachers base their comparative judgement on, a better understanding of what these Flemish teachers value in terms of argumentative texts is obtained. These insights can be further used for the professionalization aims of (pre-service) teachers or the reformulation of the attainment goals.

3. COMPARING AS A METHOD TO ELICIT WHAT TEACHERS VALUE

Within this study decision statements are used to elicit what teachers value. Decision statements are the reflections teachers provide after comparing two texts, explaining or justifying their choice for one text over the other. This reflection is the last phase of their decision making process, which was holistic and comparative in nature. Within a holistic evaluation, teachers use their own conceptualization or personal constructs to interpret and evaluate texts (Huot, 1993; Sadler, 2009a). This conceptualization comprises multiple aspects of text quality that the teachers consider valid components of the construct (Hamp-Lyons, 1987), and which develop through various kinds of experience, as for example the training they had obtained (Pula & Huot, 1993) or the community they are part of (Skar & Jølle, 2017). Moreover, this conceptualization exists in terms of more latent and more manifest aspects which teachers use more unconsciously and more consciously respectively while evaluating the texts. The manifest aspects are at the forefront, whereas the latent aspects will only be “unpacked” when the quality of a text is typical, outstanding or striking with regard to that aspect (Sadler, 2009a). So, by requesting individual teachers to provide decision statements straight after making a decision, the aspects most prominent in the teacher’s mind are elicited.

Given that decision statements result from holistic comparisons, the focus of assessors is assumed to be sharpened (Pollitt & Murray, 1996). In addition, the comparative approach might make it easier for teachers to formulate what defines quality for them (see, for example, methods such as the Kelly Repertory Grid which explicitly uses comparing and contrasting to elicit personal constructs (Kelly, 1955)). To our knowledge, only two studies have used decision statements to elicit what assessors value in essays, namely those of Whitehouse (2012) and van Daal, Lesterhuis, Coertjens, Donche & De Maeyer (2016). Whitehouse studied the decision statements that 23 geography teachers provided when comparing 564 geography essays. She concluded that the decision statements were highly related to guidelines provided for the national exams in England. Based on these findings, Whitehouse stated that teachers should have been trained extensively in the use of national guidelines and rubrics in order to be able to compare essays on relevant aspects. Next, in her study, Whitehouse provided an argument for using decision statements to investigate the validity of the ratings that result from multiple comparative judgements.

Van Daal and colleagues (2016) also used decision statements to investigate the validity of ratings obtained with regard to comparative judgments in terms of academic writing. They found that assessors use their expertise when it comes to reflecting upon their comparative judgements. In this study, 11 researchers assessed 41 scientific papers, written by students in a premaster program for a course on academic writing. A feature of this study was the selection of researchers as assessors, because they were not experienced in using the competence description for this specific course. This probably explains why the assessors discriminated between two texts based on aspects not written in compliance with the competence description.

The studies of Whitehouse (2012) and Van Daal and colleagues (2016) showed that decision statements are a powerful tool in terms of obtaining insight into the aspects that teachers value most when comparing argumentative texts. However, these studies lacked an in-depth elaboration on how teachers use their own conceptualization of text quality when comparing texts. Moreover, no information has been given on the extent to which higher or lower order aspects have informed the comparative judgements.

4. AIMS AND RESEARCH QUESTIONS

This study aims to determine which aspects are important for teachers when deciding that one text is of higher quality than another. This is important for claiming validity in terms of ratings obtained with comparative methods. Comparative methods are increasingly being used to assess texts. One of these methods is comparative judgement, wherein teachers have to compare texts with one another. These comparative methods seem to be easier and more promising in terms of obtaining reliable scores (Bramley, 2007; Steedle & Ferrara, 2016). However, studies elaborating on the aspects that the teachers take into account when comparing texts are scarce.

In this study, we will describe the aspects of argumentative texts that Flemish teachers reflect upon when explaining their comparison decisions. By relating these aspects to the final attainment goals in the Flemish context, it will be revealed how teachers use these final attainment goals in explaining their decisions. Previous studies have shown that teachers might mention the goals literally, but also change the wording to explain the components of the competence description (Wolfe, 2006). The latter reflects the use of experience in order to give meaning to the final attainment goals. In addition, other aspects might be considered by the teachers than are mentioned in the attainment goals, because they deem them to be relevant in the assessment of these specific texts (Lumley, 2002; van Daal et al., 2016). As such, insight into teachers' use of the final attainment goals will reflect upon the relevance of these goals in practice. This following question is put forward:

1) Which aspects are mentioned by teachers when reflecting upon their comparative judgements, and how do these aspects relate to the competence as defined in the final attainment goals?

Moreover, those aspects valued mostly when comparing texts will be explored. Firstly, previous research has indicated that in the case of absolute holistic scoring, the content and organization of texts attracts most attention on the part of assessors (Barkaoui, 2010; Huot, 1993; Wolfe, 2006). Within the context of academic writing, van Daal and colleagues (2016) showed that, in particular, the problem statement and the extent that the sources were analysed and synthesized informed comparative judgements. Whether teachers using a comparative approach value the same aspects is, however, unclear. Therefore, the following question is raised concerning the aspects that teachers value:

2) To what extent do teachers value the components of the construct of argumentative writing differently when using a comparative approach?

It is unclear whether teachers inform their decision by the more complex, higher order aspects or the more rule-applying lower order aspects in terms of text quality when making comparison decisions. However, it is assumed that the comparative approach is especially valuable when it comes to assessing complex competences (Pollitt, 2012a). The question is:

3) To what extent are teachers' comparative judgements based on complex and/or rule-applying components of writing?

5. METHOD

5.1 Text writers and materials

The 135 test takers, selected from ten classes within ten schools, were students in the fifth year of an 'Economics and Modern Languages' track in general secondary education in Flanders. By selecting classes from ten different schools, the participation of the schools became relevant, as we were able to provide them with feedback as to how their school performed in comparison to other schools. This kind of school feedback is scarce in Flanders. For the same reason, we used all texts of all students within this study. Students had 25 minutes to complete the writing task which was executed on school computers or laptops. During the writing task at least two researchers were present to ensure that a standardized procedure was followed. All students signed an informed consent, knowing that their texts would be anonymously used for research aims only. Their schools did not receive information on how well individual students had performed, only on how their class performed compared to the other schools.

The selection of the writing assignment was based on the following criteria: (1) matched the competence description; (2) had been successfully used in earlier scientific studies; (3) was suitable for Flemish students in the fifth year of general secondary education; (4) could be written in a short timeframe; and (5) would result in a short text. We used a previously developed and empirically tested task of van Weijen (2009) and Tillema (2012). The task was adapted to the Flemish context and was successfully pilot tested on five students. The task can be found in Appendix A.

5.2 Participants

For this study, 27 teachers participated. These teachers were reached via the university network, personal networks and websites for job-searchers. Only teachers who were working or had worked in secondary education, and therefore knew the Flemish attainment goals, were considered. Eight teachers were men, 19 women, and their average age was 45.19 (*SD* = 13.25). All of them were native speakers of the

Dutch language. Their years of relevant experience in assessing writing (mostly in a teaching context) varied from two up to 38 years (*M* = 19.96, *SD* = 13.00).

5.3 Procedure

The teachers were invited to the university campus for two afternoons. They started with an introduction to the competence description as formulated in the Flemish final attainment goals, the writing assignment to be used in this study, and the method of comparative judgement. All of them signed an informed consent, knowing that the data would be anonymously used for research aims, and that they had the opportunity to withdraw their participation at any moment. Afterwards, the teachers started with the assessment of the texts, using the D-PAC tool (www.d-pac.be), a digital tool supporting the use of comparative judgements. This tool was selected because it supports the random pairing of texts, and the random distribution of the text pairs to the teachers. Each teacher was requested to compare 23 pairs of blinded student texts. However, due to technical issues, one teacher only made 14 and another teacher 18 comparisons. In total, 606 comparisons were made. After each comparison the teachers were asked to reflect upon their decision by answering the following question: “Can you briefly explain your decision?”.

5.4 Data analysis

For 606 comparisons, teachers provided 596 decision statements. The other ten comparisons were not explained in a decision statement, as we did not force teachers to answer the question. In the first step, all decision statements were segmented into single arguments, because in most cases the teachers reflected upon more than one aspect of text quality in order to explain their decision. A segment contains a meaningful part (Braun & Clarke, 2006), explaining why a teacher chose a certain text as having a higher quality. Segments that did not provide information on an aspect of text quality were not further taken into account. These segments were, for example, referring to problems with the keyboard or the method in general. This resulted in 2054 segments for further analysis. For research question 1, the unit of analysis were these segments, whereas for research questions 2 and 3, the decision statements were the unit of analysis, with the information on whether or not an aspect of text quality is mentioned in this decision statement.

For the first research question the usual inductive content analysis was followed in order to analyse the segments. In this analysis the focus was on the aspects of text quality that were relevant to the teachers. We stuck as closely as possible to the words used by the teachers. The way teachers justified their decisions differed greatly. Some used only single words and bullet points, whereas others wrote extensive sentences and were highly descriptive in explaining why they chose a certain text over the other. This resulted in a coding scheme with different levels of abstraction. In order to ensure that the coding scheme was manageable and usable, some codes needed to be taken together, as for example spelling and grammar or capitalization and punctuation, as these aspects were most frequently mentioned together. The English version of the coding scheme, including example statements, can be found in Appendix B. Appendix C gives insight into the hierarchical structure of the codes. In order to investigate the interrater reliability of the coding process, two researchers independently coded 100 decision statements. This resulted in a Kappa of .73 across all codes, which is substantial in qualitative research (Stemler, 2001), especially considering the high number of codes analysed. Subsequently, the coding scheme was adjusted slightly. A second round was double-coded with a Kappa of .65. The lower Kappa value is probably due to the new coded segments involving some new aspects of text quality. Nevertheless, a Kappa of .65 is still substantial. The whole coding process and calculation of the Kappa were executed using NVivo qualitative data analysis software, Version 10 (2012). The codes and the number of times these codes were found in the segments will be presented for this first research question. In order to group these along with the final attainment goals, we used the coding scheme of Cumming et al. (2001, p. 26). However, we adjusted this scheme slightly because the main components differed from the final attainment goals (e.g., source use and referencing are specific for the final attainment goals but do not occur as main components in Cumming et al.'s coding scheme). An overview of the grouping can be found in Appendix C. To indicate the relationship with the final attainment goals, it will be indicated which aspect (1) referred to the final attainment goals literally; (2) referred to the final attainment goals but used other words or gave more explanation; (3) or did not directly refer to the final attainment goals. It should be noted that due to translation from Dutch into English, nuances might be lost, and meanings might be slightly changed (the Dutch coding scheme can be obtained upon request to the first author).

The second research question investigates if and how different sub-aspects of the final attainment goals have been more prominent in the decision statements. To answer this question, the codes relating to research question 1 were aggregated along the aspects of the final attainment goals. Codes not related to the goals (e.g. general statements) were not taken into account for this research question. An overview is given of the relative attention the group of teachers paid to the different component of the final attainment goals. That means that each decision statement is coded, whether an aspect was mentioned or not.

The third research question looks at the extent the group of teachers focused on complex, higher order aspects compared with rule-applying, lower order aspects of text quality. Therefore, the aspects of research question 1 are aggregated along the guidelines of Cummings and colleagues (2001, 2002). Studying the process of decision making when judging the quality of single texts, Cumming and colleagues distinguished higher order complex aspects related to the composition (rhetorical and ideational focus), as organization, style and use of sources, from the lower order more rule-applying aspects of writing (language focus), which concerns the grammar, sentences and production (e.g., length and layout). The results will present the percentages of decision statements that refer to higher-, lower-, both, or none of these components.

6. RESULTS

6.1 Aspects of text quality and their relationship with the final attainment goals

This section deals with research question 1, the aspects that segments refer to, and how these relate to the final attainment goals. Table 1 shows how some aspects that were explicitly mentioned in terms of the attainment goals were never mentioned literally in the decision statements (judgmental, integrated, content- and functional relationships, formal and content conventions of language). But 19.11% of the segments reflected literally on other aspects of the formulated attainment goals. Within the segments, many closely-related aspects were mentioned by the teachers, as for example style and the use of sources. Moreover, the aspects reflected upon by teachers show a close relationship to the final attainment goals. However, the teachers used another word (e.g., structure instead of development or references instead of citations). In other examples, teachers elaborated on an aspect of the final attainment goals e.g., stating a claim and underpinning this opinion instead of argumentation). These aspects related to the attainment goals covered 74.47% of the segments. In addition, some aspects were not part of the final attainment goals. These were more general dealing with content or form, or reflecting on (possible) characteristics of the writer, and made up 6.42% of the segments.

Table 1 Exact wording (bold), alternative wording (normal) and other aspects (italic) teachers referred to (N=2054)

Aspect	How mentioned in decision statement	*n*
Argumentation	**Argumentation**	**135**
	Judgmental	**0**
	Claim	58
	Support	92
	Elaboration	72
	Convincing	28
	Relevance	34
	Content introduction	39
	Content conclusion	56
	Content specific	62

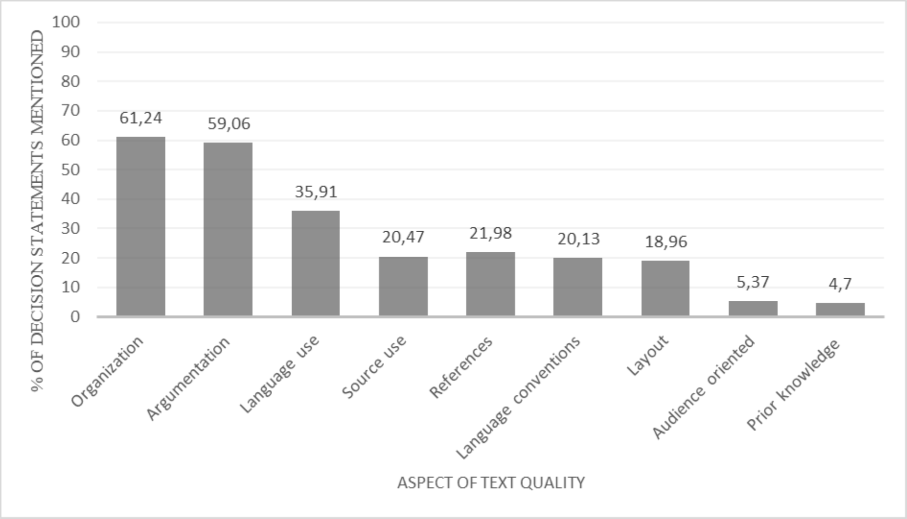
Organisation	**Development**	176
	Structure	156
	Paragraphing	44
	Coherence	77
	Content and functional relationships	**0**
	Form- and content conventions of language	**0**
Language use	Style	78
	Language and word use	77
	How it is written/ Fluency	32
	Tone	6
	Language general	73
Language conventions	Punctuation/ Capitalization	12
	Spelling/Grammar	92
	Sentence construction	48
Source use	**Integrated**	**0**
	Using and ordering information	**11**
	Use sources	113
	Integration sources	14
References	**Citing**	**36**
	Referencing	102
Audience oriented	**Orientation audience**	**11**
	Involvement of readers	12
	Focus on readers	11
Prior knowledge	**Use of prior knowledge**	**17**
	Own contribution	11
Layout	**Layout**	**4**
	Length	33
	Title present	18
	Outline and white spaces etc.	67
	Font	3
Others/ generic	*Originality*	*3*
	General	*16*
	Content general	*38*
	Formal general	*39*
	Task fulfilment	*24*
	Writer characteristics	*11*

6.2 Aspects of the final attainment goals that teachers valued most

Regarding the aspects of the final attainment goals most commonly mentioned in the decision statements, the aspects were aggregated per component for research question 2. On average a decision statement covered 2.44 (*SD* = 1.35) of the nine aspects. Figure 1 shows that teachers referred mostly to the organization (61.24% of the decision statements) of the text, with a close second best for argumentation

(59.06%). Striking is the small number of times that teachers reflected upon the extent to which a text was audience-oriented, and the use of prior knowledge, mentioned in respectively 5.37% and 4.70% of the decision statements.

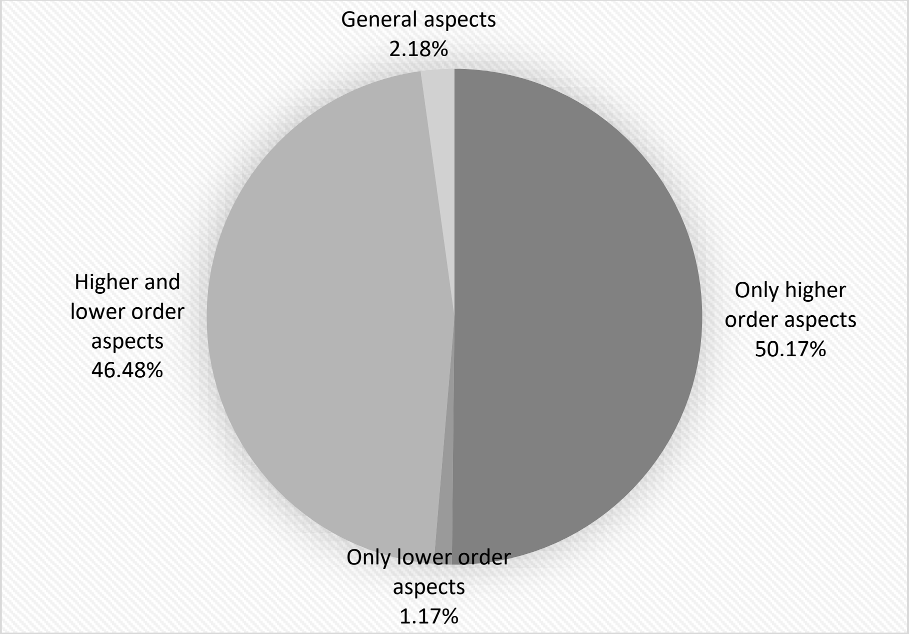
Figure 1. Percentage of decision statements referring to components of the final attainment goals (N=596)



6.3 Decision statements reflecting upon complex higher order aspects or rule-applying lower order aspects

For research question 3, the focus is on the percentage of decision statements that reflect on both, higher, or lower aspects or on general aspects. The percentage of decision statements that reflected upon complex, higher order aspects is 96.64 % (*N* = 596). As presented in Appendix C, higher order aspects cover the argumentation, organization, language use, source use, prior knowledge and audience-oriented writing. The percentage of decision statements in which teachers mention rule-applying, lower order aspects of language conventions, referencing and layout, is 47.65%. Moreover, in 46.48% complex as well as rule-applying aspects have informed teachers' comparison decisions, whereas in 50.17% only higher order complex aspects are mentioned. In only a few statements were rule-applying aspects only mentioned, namely in 1.17%. The remaining 2.18% refers to general aspects which could not be directly related to complex or rule-applying aspects. All of this is presented in Figure 2.

Figure 2. Percentage of decision statements referring to higher order, lower order both or general aspects of text quality (N=596)



7. DISCUSSION

This study reports on the aspects of argumentative texts that informed teachers' decisions within the context of comparative judgement. Therefore, 27 Flemish teachers gave an explanation for choosing one text as being of a higher quality than the other text within a text pair. This resulted in 596 decision statements. On the one hand, an in-depth analysis of these statements revealed the aspects teachers valued within argumentative writing, and how these relate to the final Flanders' attainment goals. On the other hand, overviews were given on the amount of decision statements that reflected upon different components of the final attainment goals, and the extent to which complex versus rule-applying aspects had informed teacher's comparison decisions.

Firstly, analysing all segments showed that teachers considered a wide spectrum of aspects when comparing texts. This indicates that the considered construct of text quality in a comparative method is multidimensional. Only a few of the mentioned aspects were literal reflections of the final attainment goals, which means that teachers used their experience to give meaning to these goals when comparing texts. This is in contrast to the study of Wolfe (2006), which showed that in particular the expert assessors used the wordings of the guidelines more literally in an absolute scoring procedure. A first explanation might be that the current final attainment goals are not formulated in such a way that teachers find them useful in their daily practice. For example, teachers might be used to reacting to student texts using other wordings, because these are better understood by their students. A second explanation might be that in Wolfe's (2006) study the guidelines were translated into rubrics which the assessors used regularly. Using guidelines explicitly requires regular training (Skar & Jølle, 2017). Although a lack of a thorough internalisation of the national guidelines can be a threat towards the validity of comparative methods (Whitehouse, 2012), this study evidences the contrary, as all aspects mentioned by the teachers are closely related to the goals.

Secondly, by studying which components of the final attainment goals were mentioned in the decision statements, it became clear that not all components were awarded equal attention. On the one hand, organisation and argumentation were reflected upon most frequently. This complies with studies in absolute holistic scoring that also found that these components are prominent in the minds of assessors when assessing text quality (Barkaoui, 2010; Huot, 1993). However, in contrast to van Daal and colleagues (2016), the teachers in this assessment paid less attention to source use. A possible reason is the difference in context. The text writers in van Daal and colleagues (2016) were students in a pre-master program, whereas this study reported on students in secondary education. Consequently, the requirements of the texts differed and/or the texts showed other characteristics which can be reflected in the aspects that the teachers paid attention to (Cumming et al., 2005). Another reason might be the difference in background of the assessors (researchers versus teachers in this study). Various studies elaborated on the role of the professional background in terms of the aspects that assessors consider (Eckes, 2012; Johnson & Lim, 2009; Pula & Huot, 1993). More research should be done on the impact of experience in this regard. On the other hand, audience-oriented writing and the use of prior knowledge were only in 2% of decision statements mentioned. The former can be due to audience-oriented writing being interrelated with several other components of writing, of which style and language use are the most important (Ryder et al., 1999). This relates to the multidimensionality and interrelationship between aspects of text quality that determine the construct of text quality. The reason that prior knowledge is less referred to, might be because it is unclear what it means: does it refer to, among others, knowledge about the topic or grammar or referencing rules? This ambiguity might have resulted in this aspect being coded under other codes. The findings of this study can also be interpreted as being that teachers in secondary education need more training when it comes to assessing students in terms of the final attainment goals, especially with regard to audience oriented writing and the use of prior knowledge. This is important to consider, because teachers within Flanders have a high degree of autonomy in deciding whether or not students are ready to enter the next educational level, and the formulated attainment goals should ensure that all teachers aim for these goals. Nevertheless, we can conclude that when using a holistic comparative method, teachers are encouraged to consider all components of the goals when making a decision about text quality. This means that comparative methods such as comparative judgement, are a valid way to assess text quality, as it is these decisions that inform the ratings that texts will get. In other words, when the results showed that decisions were solely based on one component of the goals (e.g., organisation), we had to conclude that comparative methods might suffer from construct underrepresentation.

Thirdly, this study also showed how complex higher order aspects of text quality were more salient to the teachers when making relative decisions on texts' quality, in contrast to the more rule-applying lower order aspects. Although the latter were mentioned in almost half of the decision statements, this was almost always accompanied with a reflection on the complex aspects. Based on this study we can assume that comparative methods in particular are a valid method for the assessment of complex skills, as the comparative methods enable the teacher to obtain reliable scores for complex skills more easily than using analytic methods (Coertjens, Lesterhuis, Verhavert, Van Gasse, & De Maeyer, 2017; Jones & Inglis, 2015) and teachers focus on these higher order skills while assessing the texts. This also means that when rule-applying skills need to be assessed, other scoring methods such as rubrics or thick boxes might be more suitable. Further research should investigate this assumption by contrasting the comparative method with other methods. For example, Barkaoui (2010), found that assessors focus on different aspects when applying absolute holistic and analytic scoring procedures using think-aloud protocols. Using an experimental approach wherein a set of texts are assessed using different rating procedures, the work of Barkaoui (2010) can be extended with insight into where assessors focus more or less when using a comparative rating procedure.

This study has limitations and shows opportunities with regard to the chosen context and the method used. Choosing argumentative writing in the context of Flanders has the consequence that the findings on the aspects mentioned above are highly dependent on this context. In other words, more research should be done on the extent to which the conclusions of this study can be replicated when other texts are assessed (Yang, Lu, & Weigle, 2015) within another country or using another type of assessor. Meanwhile, this study gives a thorough insight into how argumentative writing is understood by the teachers in Flanders. As researchers such as Osborne and Walker (2014) and Shay (2005) argue, the meaning of text quality is highly context-based, and in order to claim valid assessment, the locality is central.

Another limitation of this study is that the method is solely based on the analysis of decision statements. Decision statements revealed the aspects that teachers reflected upon, but not how these relate to the texts that the teachers chose. Moreover, some decision statements were very elaborated while others only contained one word. It is unclear whether this mirrors differences in how teachers make decisions (see for example Vaughan, 1991), or whether this is due to the elicitation method used. It would be useful to examine how decision statements relate to the aspects that text quality assessors pay attention to when reading and evaluating both texts. Using a think aloud approach would enrich our understanding as to how teachers execute comparative judgements. An advantage of decision statements is, however, that they allow us to analyse the aspects that multiple teachers value when making several comparative judgements on a detailed level. This opens opportunities for further research. For example, when multiple comparisons of one teacher are to be analysed, decision statements offer an interesting approach to elicit teacher's conceptualizations of text quality.

To conclude, this study shows which aspects of argumentative writing are valued by teachers while comparing argumentative texts. Besides the already argued reliability and efficiency of the comparative approach (Coertjens et al., 2017; Steedle & Ferrara, 2016), the variety of the aspects mentioned, the focus on organisation and argumentation, and more complex aspects, seem to be promising for claiming that comparative judgements can lead to valid ratings.

AUTHORS' NOTE

This work originates from project 'D-PAC: Development, validation and effects of a Digital Platform for the Assessment of Competences' and was supported by the Flanders Innovation & Entrepreneurship [grant number 130043].

REFERENCES

Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice: Developing language assessments and justifying their use in the real world* (Vol. 1). Oxford, UK: Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83–100. https://doi.org/https://doi.org/10.1177/0265532214542994

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. A. Baird, H. Goldsteing, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 294 - 300). London, UK: QCA.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, G. T. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly, 64*(3), 276–291. https://doi.org/10.1111/j.1468-2273.2010.00460.x

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly, 12*(3), 262–282. https://doi.org/10.1080/15434303.2015.1053134

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London, UK: Routledge.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering [Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment]. *Pedagogische Studien: Tijdschrift Voor Onderwijskunde En Opvoedkunde*, *94*(4), 283–303

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. Ewing, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series, 1961*(2), i-93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice, 20*(3), 308–324. https://doi.org/10.1080/0969594X.2013.779229

Hamp-Lyons, E. M. (1987). *Testing second language writing in academic settings*. University of Edinburgh, Edinburgh.

Heldsinger, S. A., & Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1–19. https://doi.org/10.1007/BF03216919

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing, 36*, 69–74. https://doi.org/10.1016/j.jslw.2017.05.002

Huot, B. A. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press, Inc.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485–505. https://doi.org/10.1177/0265532209340186

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kelly, G. (1955). *Personal construct psychology*. New York, NY: Norton.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative judgement as a promising alternative to score competences. In G. Ion & E. Cano (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 120-140). Hershey, PA: IGI Global.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research, 64*(4), 147–156.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13-103). New York, NY: Macmillan Publishing.

Messick, S. (1994). *Alternative modes of assessment, uniform standards of validity* (No. 2) (p. 1–22). ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.1994.tb01634.x

Moss, P. A. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing, 1*(1), 109–128. https://doi.org/10.1016/1075-2935(94)90007-8

Osborne, J., & Walker, P. (2014). Just ask teachers: Building expertise, trusting subjectivity, and valuing difference in writing assessment. *Assessing Writing, 22*, 33–47. https://doi.org/10.1016/j.asw.2014.06.002

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Pollitt, A., & Crisp, V. (2004). Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions? Presented at the British Educational Research Association Annual Conference, Manchester. Retrieved from: http://www.leeds.ac.uk/educol/documents/00003731.htm

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (Vol. 3), (pp. 74–91). Cambridge, UK: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Preiss, D. D., Castillo, J. C., Grigorenko, E. L., & Manzi, J. (2013). Argumentative writing and academic achievement: A longitudinal study. *Learning and Individual Differences, 28*, 204–211. https://doi.org/10.1016/j.lindif.2012.12.013

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Creskill, NJ: Hampton Press, Inc.

Ryder, P. M., Vander Lei, E., & Roen, D. H. (1999). Audience considerations for evaluating writing. In R. Cooper & L. Odell (Eds.), *Evaluating writing: the role of teachers' knowledge about text, learning, and culture* (pp. 53–71). Newark, IL: National Council of Teachers of English.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sadler, D. R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Sadler, D. R. (2009b). Transforming holistic assessment and grading into a vehicle for complex learning. In In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 1–19). Dordrecht, The Netherlands: Springer.

Sakyi, A. A. (2003). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. University of Toronto, Toronto.

Schipolowski, S., & Boehme, K. (2016). Assessment of writing ability in secondary education: comparison of analytic and holistic scoring systems for use in large-scale assessments. L1 Educational Studies in Language and Literature, 16, 1–22. https://doi.org/10.17239/L1ESLL-2016.16.01.03

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Shay, S. (2005). The assessment of complex tasks: a double reading. *Studies in Higher Education, 30*(6), 663–679. https://doi.org/10.1080/03075070500339988

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: Investigation of a long term writing assessment program. *L1 - Educational Studies in Language and Literature, 17*, 1–30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education, 29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation, 7*(17), 137–146.

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes. Netherlands Graduate School of Linguistics, Utrecht*. Retrieved from http://www.lotpublications.nl/writing-in-first-and-second-language-writing-in-first-and-second-language-empirical-studies-on-text-quality-and-writing-processes

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 1–16. https://doi.org/10.1080/0969594X.2016.1253542

van den Bergh, H., De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27) (pp. 23–32). Leiden, The Netherlands: Koninklijke Brill NV.

Van Grinsven, V., Mondrian, L., & Westerik, H. (2007). *Taalpeil-onderzoek 2007. Onderwijs Nederlands in Nederland, Vlaanderen en Suriname*. Den Haag, The Netherlands: Nederlandse Taalunie.

van Weijen, D. (2009*). Writing processes, text quality, and task effects: Empirical studies in first and second language writing*. Netherlands Graduate School of Linguistics, Utrecht. Retrieved from http://www.lotpublications.nl/writing-processes-text-quality-and-task-effects-writing-processes-text-quality-and-task-effects-empirical-studies-in-first-and-second-language-writing

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

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

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

Weigle, S. C., & Montee, M. (2012). Rater's perceptions of textual borrowing in integrated writings tasks. In E. Van Steendam, M. Tillema, G. C. W. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 117–145). Leiden, The Netherlands: Koninklijke Brill NV.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400–409.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester, UK: AQA Centre for Education Research and Policy.

Wingate, U. (2012). "Argument!" Helping students understand what essay writing is about. *Journal of English for Academic Purposes, 11*(2), 145–154. https://doi.org/10.1016/j.jeap.2011.11.001

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*(1), 37–56.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59–75. https://doi.org/10.1080/09695940903565610

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

APPENDIX A – WRITING TASK "GETTING CHILDREN"

Having children, yes or no?

The Flemish Organization of Students is organising a national essay contest, especially for pupils in the fifth year of general education. You're also taking part. You absolutely want to win. The winning essay will be printed in Yeti, a monthly magazine that is read by pupils your age from all over Flanders.

The subject of the essay has already been decided and was described in Yeti as follows:
Until recently people usually assumed that marriage should, in principle, lead to parenthood. Over recent years, however, the number of people who decide not to have children has strongly increased. The question "children, yes or no" is now much more an issue than it used to be, and having children is not that obvious any more. Apart from that, it is sometimes thought that the answer to this question depends on your convictions about life, and the ideas you have about what society should look like.

Assignment:
Write an essay in which you give your opinion to the question:
"Having children, yes or no?"

The essay has to meet the following requirements, set by the jury:

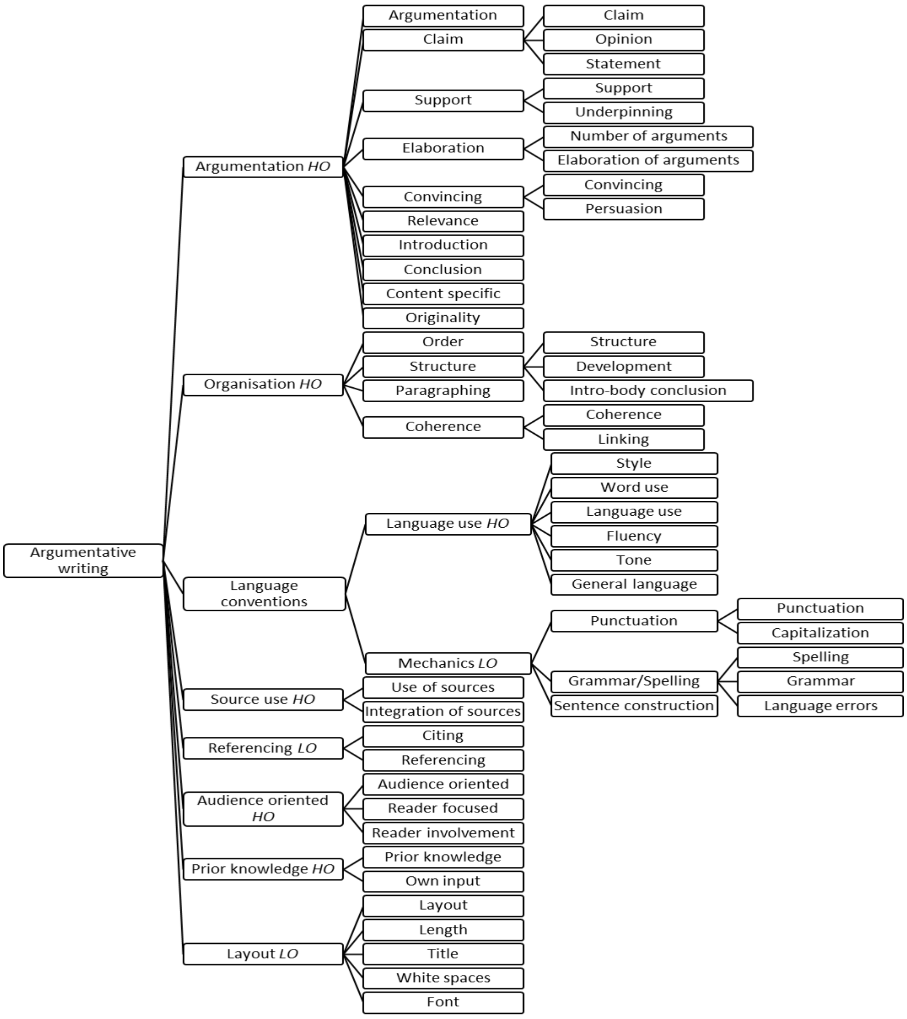
1) Your essay must be about half a page in length.
2) You must do your best to convince your readers, fellow pupils, of your opinion.
3) You must give arguments to support your opinion.
4) Your essay must be structured in a good and logical way.
5) Your essay must look well cared for (think of language use and spelling).
6) In your essay you must use at least two extracts from the 'References' (can be requested by the first author). You must include these extracts in your essay in a meaningful way.

You have 25 minutes to complete this assignment.
Good luck!

APPENDIX B – CODING SCHEME WITH EXAMPLE STATEMENTS

Content specific	However, the idea behind the comparison between two couples and how children fit in their lives is good	231
Convincing	The right text really tries to convince	1591
Development	The right texts develops well	109
Elaboration	Not every argument is elaborated on (especially at the end of the text)	333
Fluency	The text is easy readable	545
Focus on readers	The audience is literally spoken to	78
Font	Also, with regard to formal correctness (font, ….), this author scores better	1505
Form general	The form aspects are used in the right text	2044
General	I think neither is good	625
Grammar and spelling	The right text contains more grammar and spelling mistakes	60
Introduction	The introduction is better	2066
Involvement readers	Text 2 involves the audience perfectly	2181
Language general	However, the right text is better due to language and …	83
Language use	There is a beautiful and creative use of language	149
Layout	In the left text, more attention has been paid to layout	60
Length	The left text is too short	159
Ordering information	Information has been …. and ordered	149
Orientation audience	Is not oriented to the unknown public	93
Originality	The author is the first with an original perspective	379
Paragraphing	The left text has no paragraphs, which makes is more difficult to read	199
Prior knowledge	An obvious use of available prior knowledge	284
Punctuation	The second text is weak regarding…, punctuation, …	441
Referencing	They both refer to sources	292
Relevance	In the right text, most arguments are not related to the topic	166
Sentences	Left: already at the start it is obvious that the author does not use sentences but just puts words together without any thought being given to sentence construction	1640
Source integration	The sources are processed in the text	37
Source use	The text uses more fragments from sources	397
Structure	The left text has a better structure	114
Style	However, the right text is better due to … and style	83
Support	It is a bit better supported	305
Task fulfillment	The text better complies with the question asked	95
Title	The text has a title	1814
Tone	The tone is too bombastic	389
White spaces	White spaces in the right text makes it easier to read	44
Word use	It has a bad word choice	515
Writer characteristic	The author of the right text is obviously too young to write about this theme	379

APPENDIX C – ORDERING OF THE CODES (HO = COMPLEX, HIGHER ORDER ASPECTS; LO = RULE APPLYING, LOWER ORDER ASPECTS)


Argumentative writing
Argumentation *HO*
Argumentation
Claim
Claim
Opinion
Statement
Support
Support
Underpinning
Elaboration
Number of arguments
Elaboration of arguments
Convincing
Convincing
Persuasion
Relevance
Introduction
Conclusion
Content specific
Originality
Organisation *HO*
Order
Structure
Structure
Development
Intro-body conclusion
Paragraphing
Coherence
Coherence
Linking
Language conventions
Language use *HO*
Style
Word use
Language use
Fluency
Tone
General language
Mechanics *LO*
Punctuation
Punctuation
Capitalization
Grammar/Spelling
Spelling
Grammar
Language errors
Sentence construction
Source use *HO*
Use of sources
Integration of sources
Referencing *LO*
Citing
Referencing
Audience oriented *HO*
Audience oriented
Reader focused
Reader involvement
Prior knowledge *HO*
Prior knowledge
Own input
Layout *LO*
Layout
Length
Title
White spaces
Font