UNLOCKING GENRE KNOWLEDGE THROUGH TEXT EXEMPLAR AND ASSESSING ITS IMPACT ON ARGUMENTATIVE ESSAY QUALITY

TINE MOMBAERS, ROOS VAN GASSE, AND SVEN DE MAEYER

University of Antwerp

Abstract

This quasi-experimental study investigates the impact of analyzing text exemplars on genre knowledge and on the quality of argumentative essays in authentic educational settings. Focusing on single and analogue text exemplars, we assess whether targeted genre knowledge and text quality can be improved. Conducted in classrooms with teacher-led Socratic instruction and individual text analysis, our hypotheses explore the effectiveness of both single and analogue exemplars on genre knowledge and text quality. Results show that while both exemplar types enhance genre knowledge similarly, only analogue exemplars significantly improve text quality. The study underscores the complex relationship between genre knowledge and text quality, suggesting that factors beyond genre knowledge contribute to writing proficiency. The findings highlight the nuanced process of knowledge transfer in writing and the potential of analogue exemplars to facilitate this transfer effectively.

Keywords: learning from exemplars, comparisons, argumentative writing, genre knowledge, text quality

Mombaers, T., Van Gasse, R., & De Maeyer, S. (2025). Unlocking genre knowledge through text exemplars and assessing its impact on argumentative essay quality. *L1-Educational Studies in Language and Literature, 25, 1-41. https://doi.org/10.21248/l1esll.2025.25.1.791*

Corresponding author: Tine Mombaers, University of Antwerp, Sint-Jacobstraat 2, 2000 Antwerp, Belgium. Email: tine.mombaers@uantwerp.be

1. INTRODUCTION

In contemporary society, writing poses a significant challenge for students. They often perceive writing as a difficult and daunting activity (Hidi & Boscolo, 2006). Over the past decades, students' writing skills have proven to be poor across educational levels in the international context (e.g., Abrams, 2019; Bañales et al., 2018; Da Cunha & Montané, 2019; Johnson et al., 2017). According to Bacha (2010), students particularly struggle with argumentative writing. Students' proficiency in argumentative writing seems to be lacking both in secondary school and higher education settings (Graham & Perin, 2007; NCES, 2012; Ferretti & Lewis, 2013; Song and Ferretti, 2013). Students have difficulties with recognizing and applying argumentative text structures (Chambliss & Murphy, 2010; Freedman & Pringle, 1984), generating evidence (Kuhn, 1991), offering relevant reasons (McCann, 1989) and producing counterarguments (e.g., Perkins et al., 1991; Stapleton, 2001; Nussbaum & Kardash, 2005). The lack of integrating counterarguments leads to poorly developed arguments that fail to consider alternative viewpoints. This tendency to consider only the side of the issue favored by the student was labeled as my-side bias by Perkins et al. (1991).

This below standard performance in argumentative writing poses a societal concern, given the importance of strong argumentative skills in academic, personal, and professional contexts (Lee & Deakin, 2016; Pessoa et al., 2017). After all, proficiency in argumentative writing, which contains the skills to analyze, compose, and evaluate well-founded arguments, is crucial for academic success (Muller Mirza & Perret-Clermont, 2009; Newell et al., 2011).

Furthermore, good argumentative writing skills nurture critical thinking skills (Kuhn & Crowell, 2011), empowers individuals to influence others, and it fosters debating controversial issues enabling greater participation in social and democratic processes (Ferretti et al., 2009). Additionally, proficiency in argumentative writing is essential to be prepared for the modern workplace (Ferretti & De La Paz, 2011).

Considering the challenges that students face in argumentative writing, it is paramount for researchers to explore effective ways to improve students' proficiency in this area.

Numerous studies have demonstrated the teachability of argumentative writing and its potential for improvement (e.g., Granado-Peinado et al., 2019; Nussbaum & Schraw, 2007; Prata et al., 2019). Successful interventions have used various methods to enhance students' writing. On the one hand, providing peer feedback (Latifi et al., 2021), computer based formative assessments (Moschella, 2023), and automated writing evaluation (Zhai & Ma, 2023) can enhance argumentative writing. On the other hand, instructional approaches have also shown promise. Granado-Peinado et al. (2019) used explicit instruction with video modelling and collaboration with other students. Nussbaum and Schraw (2007) provided explicit instruction (alongside the use of a graphic organizer) in argument—counterargument integration, whereas Prata et al. (2019) deployed SRSD (Self-Regulated Strategy

Development) instruction combined with a cooperative method. In these intervention studies, certain aspects of genre knowledge play a part in either giving feedback (Latifi et al., 2021; Moschella, 2023), or including it in (part of) their instruction (Granado-Peinado et al., 2019; Nussbaum & Schraw, 2007). After all, genre knowledge plays an essential part in gaining argumentative writing skills. Olinghouse et al. (2015) found that genre knowledge is significantly correlated with writing quality and the use of genre elements in argumentative texts with 5th graders. These findings imply that genre knowledge can play an important role in writing qualitative argumentative texts. Several empirical studies on writing emphasize the importance of genre knowledge early on in elementary school and also in secondary school (e.g., Bigger, 2022; Olinghouse & Graham, 2009; Olinghouse et al., 2015) or incorporate increasing genre knowledge as a part of their intervention, among other things, to enhance students' writing within the elementary or university context (De Smedt & Van Keer, 2018; Yasuda, 2011)). However, none of these studies in writing exclusively focuses on enhancing genre knowledge as a means to improve students' writing. To the authors' knowledge, there is only one study that that focuses solely on enhancing genre knowledge of argumentative texts (Mombaers et al., 2024). We found in our previous study that genre knowledge can be improved through learning from text exemplars. However, there is currently no evidence indicating that the increase in genre knowledge in itself will lead to better argumentative texts.

In contrast to previous studies, which implemented comprehensive instructional designs to improve students' writing (Granado-Peinado et al., 2019; Nussbaum & Schraw, 2007; Prata et al., 2019), this study aims to explore the potential impact of enhancing a singular crucial aspect: genre knowledge. By focusing solely on this specific element within argumentative writing, we aim to determine whether targeted enhancements in genre knowledge can lead to improved text quality.

2. LITERATURE

In this section, we initiate an exploration of writing, delving into the complex domains of genre knowledge, learning from (comparing) exemplars and argumentative writing. Genre knowledge is fundamental for understanding different writing forms and facilitating effective communication across various contexts. Learning from exemplars emerges as a strong strategy for improving genre understanding. Additionally, we describe the challenges of argumentative writing and improving text quality. Through this oversight of relevant literature, we aim to provide insights into the varied aspects of writing and strategies for improving text quality.

2.1 Genre knowledge

Genre knowledge, as outlined by McCutchen (1986), entails recognizing the specific characteristics of different types of writing. This involves understanding the purpose,

content, and structure of various types of written works, such as essays, stories or reports (Martin, 2009). Additionally, Hyon (2001, 2002) suggests that genre knowledge also encompasses recognizing the language style used in different types of writing.

Beyond its core definition, genre knowledge plays a crucial role in various writing contexts. It serves as a cognitive framework that writers use when facing new writing tasks in unfamiliar contexts. This conceptual framework aids in bridging rhetorical and social understanding (Beaufort, 2007; Tardy, 2009), enabling students to recognize and adapt to new writing contexts more effectively, as highlighted by Miller (2009).

Theoretical considerations suggest that genre knowledge positively influences writing quality (e.g., Donovan & Smolkin, Gillespie et al., 2013; Olinghouse & Graham, 2009; Saddler & Graham, 2007). Students who possess a deeper understanding of basic genre elements or are more familiar with a genre, can use this knowledge to define the writing task, guide the retrieval of relevant information (such as ideas and vocabulary), and assess the appropriateness of the retrieved ideas. Consequently, students are expected to produce better-quality papers with appropriate genre elements (Olive et al., 2009; Donovan & Smolkin, 2006; Saddler & Graham, 2007).

These theories find support in empirical research. For example, Olinghouse et al. (2015) discovered that students who grasp different types of writing perform significantly better in various writing tasks, like narrative, informative and persuasive writing. Genre knowledge positively affects students' overall writing quality in this study, as well as the inclusion of genre elements in their narratives, persuasive essays, and informative texts. In their study, genre knowledge was positively and significantly correlated (0.51) with text quality. Text quality in this study was assessed through a genre-specific, holistic rubric, including the following aspects: development of ideas, organization, sentences/word choice/voice, and genre elements. Further research by Ferretti and Lewis (2019) affirmed the positive impact of genre knowledge on argumentative writing. Students with deeper understanding of persuasion and persuasive writing demonstrated the ability to produce higher-quality persuasive essays compared to those with limited genre knowledge. Moreover, studies indicate that instructing genre elements contributes positively to the overall quality of student writing (De Smedt & Van Keer, 2018; Graham et al., 2012; Koster et al., 2015).

2.1.1 Learning from (comparing) exemplars

In recent decades, numerous studies (e.g., De Smedt et al., 2018; Graham et al., 2012; Olinghouse et al., 2015) have underscored the significance of genre knowledge, highlighting the need to explore effective methods for its enhancement. In our previous study (2024), we investigated if learning from text exemplars (individually) could enhance genre knowledge, building upon the scientific

importance of genre knowledge for text quality and the limited exploration of genre knowledge as a distinct variable in literature.

In this section, we will first describe the general literature on learning from exemplars. Consecutively, we will zoom in on the literature on the use of exemplars within writing education.

The process by which individuals develop conceptual knowledge has been widely studied in cognitive psychology. Learning from exemplars has emerged as a promising method to facilitate this process (Alfieri et al., 2013; Gentner, 1983). In this approach, learners are exposed to exemplars of varying quality and subsequently assess different aspects of these exemplars in terms of their quality. Individuals tend to learn more effectively when comparing exemplars than by processing single exemplars (Alfieri et al., 2013). Gentner's (1983) structure-mapping theory provides a theoretical underpinning for this finding, emphasizing that comparing exemplars enhances the salience of common aspects and facilitates abstraction. Exemplars can be presented either individually or in pairs, with paired presentations categorized as analogue, near miss, or contrastive comparisons. So, by comparing different exemplars, students can notice their common features, potentially enhancing their conceptual knowledge. These exemplar types differ in the degree of similarity and difference between the paired exemplars, shaping how learners process and organize the information (Alfieri et al., 2013; Hammer et al., 2008; Smith & Gentner, 2014). Analogue comparisons involve exemplars that are highly similar and typically belong to the same category. Learners align shared features and relationships, which helps them identify overarching patterns and develop category abstractions or schemas (Gentner & Namy, 1999; Namy et al., 2007). Near miss comparisons feature exemplars that are similar in many aspects but differ in one key feature, which is crucial for distinguishing between categories. The overlap between these exemplars naturally draws attention to the critical difference, enabling learners to refine their understanding of category boundaries (Hammer et al., 2008). Contrastive comparisons, in contrast, pair exemplars from different categories. These exemplars share some similarities but also exhibit significant differences, requiring learners to identify both the overlapping and distinct attributes to understand how categories differ (Smith & Gentner, 2014). Thus, these three types of comparisons—analogue, near miss, and contrastive—illustrate different relationships between exemplars, aiding learners in recognizing patterns, distinguishing critical features, and understanding category contrasts.

Although the literature on learning from single exemplars is limited, existing evidence suggests that sequential presentation of exemplars can be effective for both adults and children (Childers, 2008; Reed, 1987; Ross & Kennedy, 1990). The learner's capability to align the representations presented in sequence influences the success of such learning (Christie and Gentner, 2010). However, comparison is deemed critical for relational abstraction, as emphasized by the same authors and this is backed by the meta-analysis by Alfieri et al. (2013). In this meta-analysis, the authors conclude that individuals tend to learn more effectively when comparing

exemplars than by processing single exemplars sequentially. In addition, learning from analogue exemplars outperforms other learning situations like single cases, traditional instruction, and non-analogous cases, with a medium effect size (Alfieri et al., 2013).

Hammer et al. (2008) demonstrated that learning from contrasting exemplars engages significant cognitive processes compared to learning from same-class or analogue comparisons. Furthermore, the extent of difference between the exemplars plays a role in moderating learning outcomes from contrasting exemplars. Among these, 'near miss' exemplars are identified as particularly effective (Smith & Gentner, 2014; Hammer et al., 2008). Analyzing 'near miss' exemplars (exemplars with much overlap and one key difference) involves a self-aligning comparison, wherein essential aspects become more prominent in the learning process through aligning similarities and differences between exemplars (Smith & Gentner, 2014). Additionally, students can also learn from contrastive exemplars (many differences and some overlap), according to the same authors.

Several studies (e.g., Alfieri et al., 2013; Gadgil & Nokes, 2009; Kurtz et al., 2001; Mombaers et al., 2024) suggest that learning through comparisons can be an effective method for acquiring more complex skills. In the educational contexts that were investigated, students often work independently and were assigned to compare exemplars by looking for similarities and/or differences. In the realm of writing education, particularly in higher education, text exemplars are typically utilized differently. They are frequently integrated with traditional teacher instruction (Hyland, 2007; Tribble, 2015), where teachers employ genre exemplars to highlight textual features for learners. Alternatively, students may receive high-quality exemplars as part of feedback on their work, serving as a tool to enhance their own writing (Smyth & Carless, 2020; Lipnevich, 2014; To et al., 2022). Moreover, exemplars are employed to showcase the expected quality and demonstrate the type of text that would perform well across all components of the assessment criteria used to evaluate students' work (Carless & Chan, 2017; Hendry et al., 2011).

In teaching genre knowledge, model texts cand be used to compare and contrast. However, Abbuhl (2011) and Charney and Carlson (1995) found that providing students opportunities to study these models is insufficient to enhance their writing skills. Teachers need to explicitly explain and describe the different aspects of genre and text structure in the model texts for students to effectively acquire this knowledge (Abbuhl, 2011). This is contradicted by recent findings of our previous study (2024). This quasi-experimental study, conducted with 76 students in the 11th grade in a classroom setting, involved four groups analyzing single and sequential, analogue, near miss, and contrastive text exemplars. The study's aim was to examine whether analyzing text exemplars could lead to improved genre knowledge of argumentative texts. Students actively and independently examined and compared text exemplars to identify similarities and differences. We conducted such a study, enabling students to independently compare argumentative text exemplars to

enrich their genre knowledge of argumentative texts. The results revealed that individual learning (so without teacher instruction) from single, sequential text exemplars and analogue text comparisons significantly enhances students' genre knowledge of argumentative texts. Interestingly, the study unveiled that there was no significant difference in the effectiveness of students' genre knowledge acquisition when they either examined single text exemplars sequentially or engaged in analogue comparisons. In addition, we did not observe a positive effect of comparing near miss and contrastive text exemplars on genre knowledge (2024). Two possible explanations were put forward (Mombaers et al., 2024). Firstly, making comparisons that require identifying both similarities and differences may be the least effective method to make text comparisons. Secondly, students had to compare one text of lower quality with another of higher quality because texts had to differ in the presence of genre elements. Thus, the text quality of the exemplars might have influenced the results.

2.2 Argumentative writing

Building on genre knowledge, mastering the specifics of argumentative writing is crucial for producing high-quality texts. The genre conventions of argumentative writing involve identifying a central claim, presenting supportive evidence (whether empirical or experiential), and critically evaluating warrants that establish connections between the claim, evidence, and the broader contextual situation constituting an argument (Newell et al., 2011). To achieve persuasiveness, an argumentative essay must exhibit a robust surface structure by including alternative perspectives and clarifying their limitations. Simultaneously, it should underpin its claims with compelling and high-quality reasons that convince the audience (Kuhn, 1999; Stapleton & Wu, 2015).

Toulmin's framework (1958, 2003) offers a foundational structure for effective argumentation, widely recognized and cited within the field. It isolates key elements of sound argumentation: claim (the initial conclusion), data (supporting facts), warrants (connecting data to the claim), backings (assumptions underpinning warrants), qualifiers (limitations on claim strength), and rebuttals (arguments challenging or providing exceptions to the elements of the argument).

While Toulmin's framework has been instrumental in emphasizing the need to consider alternative positions, it has faced criticism for overemphasizing structural elements at the expense of logic and evidence (Sampson & Clark, 2008). Researchers have encountered difficulties in its application, as the arguments students wrote could often be allocated to more than one element (Sampson & Clark, 2008; Simon, 2008). Concerns about its complexity led to adaptations and simplifications to ensure reliable classification of argumentative elements (Nussbaum & Kardash, 2005; Qin & Karabacak, 2010; Stapleton, 2001; Stapleton & Wu, 2015).

For example, Stapleton and Wu (2015) simplified Toulmin's framework to explore the quality of reasoning in students' essays. They developed a rubric based on a

modified Toulmin model, assessing the relationship between surface structure elements (e.g., claims and counterarguments) and the substantive quality of reasoning. Despite a strong surface structure, many claims and supporting data in students' essays were considered weak, emphasizing that effective reasoning does not always align with a well-structured surface. Consequently, they introduced the Analytic Scoring Rubric for Argumentative Writing (ASRAW), evaluating both argumentative structural elements and reasoning quality (see Appendix A). This rubric, widely used by researchers (Abdollahzadeh et al., 2017; Allagui, 2021; Mohsen & Qassem, 2020; Mombaers et al., 2024), contains specific components of argumentative writing, aiding in identifying genre-specific elements in students' writing.

3. THIS STUDY

In this quasi-experimental study conducted in an authentic educational setting, our aim is to explore the potential impact of enhancing a singular crucial aspect: genre knowledge. By focusing solely on this specific element within argumentative writing, we seek to determine whether targeted enhancements in genre knowledge through learning from text exemplars can lead to improved text quality. In the current study, we opted to exclusively employ single and analogue text exemplars since we found these two to be effective to enhance genre knowledge (Mombaers et al., 2024).

Since the study took place in authentic classrooms, teachers were responsible for carrying out the lessons and collecting most of the data. In everyday teaching, teacher instruction is an essential part of teaching practice. Therefore, we chose to include teacher instruction, utilizing Socratic methods, along with individual analysis of text exemplars. This combination aimed to enhance the learning experience by integrating guided teaching with independent analyses of text exemplars. Building on our previous study on enhancing genre knowledge through learning from text exemplars (2024), we hypothesize the following:

> H1: Genre knowledge can be improved through learning from single text exemplars.
>
> H2: Genre knowledge can be improved through learning from analogue text exemplars.
>
> H3: Analyzing either single or analogue text exemplars will result in similar levels of improvement of genre knowledge.

Furthermore, we argue that enhancing genre knowledge of argumentative texts through learning from text exemplars with teacher instruction and individual work can lead to improved text quality. Therefore, we hypothesize that *Text quality can be improved through analyzing single and analogue text exemplars (H4)*.

Additionally, we hypothesize that there will be no significant difference in the enhancement of text quality between the single and analogue condition, based on the finding that there is no differential effect in learning from single and analogue exemplars (Mombaers et al., 2024). Hence, the hypothesis is: *Analyzing single or*

analogue text exemplars will not lead to differential improvements in text quality (H5).

Finally, given the positive impact of genre knowledge on students' writing performance (Ferretti & Lewis, 2019; Olinghouse et al., 2015), we also anticipate a correlation between genre knowledge and text quality. *Genre knowledge is correlated with text quality (H6).*

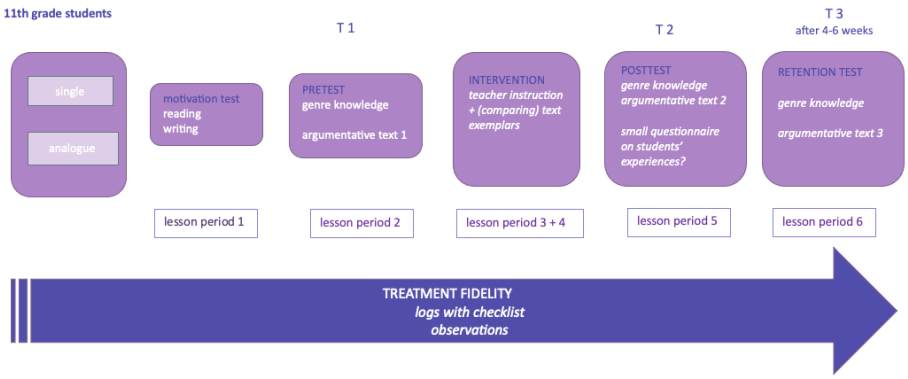
We intentionally use the term 'text quality' which, in this study, is operationalized by the extent to which all essential genre elements of argumentative essays are included (ASRAW). We will not deploy the terms 'writing proficiency', 'writing skills', nor 'writing quality', since we believe that these concepts cover more than text quality. Quality of writing can be evaluated considering among others syntactic complexity, the use of specialized vocabulary, text cohesion, text length, word frequency, sentence length, grammatical structures, writing planning, and metacognitive knowledge (Beauvais et al., 2011; Beers & Nagy, 2009; Crossley & McNamara, 2016; Dikli, 2006; Guo et al., 2013; McNamara et al., 2010).

4. METHODS

To answer the research questions above, an intervention study was set up with two conditions (single and analogue) in an authentic classroom setting in secondary education. Teachers received a complete lesson package that entailed all lessons for pretest, intervention, posttest and retention test. These materials can be found on OSF (https://osf.io/3ht7e/?view_only=6da7d13f961046c99718d62463dee822).

Data was gathered during school hours in the fall of 2023 (September till November).

Figure 1. Research design of the intervention study



4.1 Participants and setting

The desirable number of participants was calculated through power analysis, using G*Power (Faul et al., 2007). To achieve a power of 0.80 with an effect size of 0.20, two groups, and three measurements, a sample size of 42 participants was needed. This sample size was then adjusted, considering the nested structure of the data (Hox et al., 2017), resulting in a final desirable sample size of 82.

The study involved students from 18 class groups, from five secondary schools located in Flanders, the Dutch-speaking region of Belgium. 158 students from various fields of study participated in the intervention study, as shown in table 1.

Of these participants, 58.86 % were girls and 41.14 % were boys. Their ages ranged from 15 to 18.

Table 1. Overview of study participants and their field of study

School	Class group / study program	Focus of study program	Condition	N students
1	Human Sciences	THE cross-domain	single	20
1	Greek & Latin / Greek & Math / Latin & Modern Languages / Latin & Sciences / Latin & Math	THE cross-domain	single	15
1	Modern Languages / Modern Languages & Sciences / Latin & Modern Languages	THE cross-domain	single	8
2	Business & organization	THE cross-domain	single	12
2	Education & guidance	THE & LM	single	15
2	Welfare Sciences	THE domain-specific	analogue	11
2	Latin & Sciences / Sciences & Math	THE cross-domain	analogue	4
2	Economics & Modern Languages	THE cross-domain	analogue	7
3	Languages & communication	THE & LM	analogue	5
3	Application and data management	THE & LM	analogue	9
3	Commercial organization	THE & LM	analogue	9
4	Economics & Modern Languages / Modern Languages & Sciences / Latin & Modern Languages / Languages & communication	THE cross-domain	single	11
4	Welfare Sciences & Business Sciences	THE domain-specific	analogue	5
4	Sciences & Math	THE cross-domain	analogue	14

5	Computer & Communication Sciences	THE domain-specific	single	1
5	Biotechnical and Chemical STEM Sciences /	THE domain-specific	analogue	6
5	Construction and carpentry Sciences	THE domain-specific	analogue	
5	Technological Sciences & Engineering	THE domain-specific	single	6

Note. THE = Transition into Higher Education, THE & LM = Transition into Higher Education or Labor Market

Class groups were assigned to the conditions through stratified sampling. We divided students into the two conditions, making sure that in each condition included a variation of (focus of) study programs. Students in this sample were enrolled in either a study program focused solely on preparing them for higher education (THE), or a program with a dual focus, offering prospects to both higher education and the labor market (THE & LM). We had an original sample of 228 students, with 116 students in the single condition and 112 students in the analogue condition. Not all students participated in the study due to consent issues, which led to a total of 158 students divided into two conditions: single (N=88) and analogue (N=70).

4.2 Teaching training

Since the teachers involved in the study were responsible for the data collection and the lesson series, they received a two-hour online training at the beginning of September. They received information on how to gather student and teacher consent, the content and planning of the lessons, and background information about the research project. All teachers were given an extensive teacher's manual in which they found the intervention protocol to which they had to comply (timing, do's and don'ts, etc.). The teacher's manual can be found in the materials section on OSF (https://osf.io/3ht7e/?view_only=6da7d13f961046c99718d62463dee822).

A Teams environment was set up to share all files with the teachers and they could also use this platform to ask questions.

4.3 Intervention

The intervention comprised of a lesson series of two lesson periods (100 minutes in total). In the first lesson period, students received instruction on genre elements of argumentative essays, through mostly the Socratic method. Teachers discussed with their students what genre knowledge is, also making the distinction between specific and more general genre knowledge. They worked with one (single condition) or two (analogue condition) text exemplars to look for genre elements of argumentative essays. Next to genre knowledge, attention was paid to specific ways to express one's opinion and signaling words that are appropriate in argumentative essays. In the second lesson period, students received individual assignments in which they

had to look at and analyze six text exemplars. In the single condition, the text exemplars were sequentially presented to the students. They were asked: *"Which genre elements of an argumentative essay are present in the texts?"* Students in the analogue condition were presented with the text exemplars in three pairs. They had to answer the question *"What similarities do you see in both text 1 and 2?"* This question was selected because previous research proves that letting students focus on similarities in analogue exemplars is most effective (Alfieri et al., 2013). Students in both conditions were given a minimal number of genre elements that they had to be able to find. In both conditions, students were explicitly asked to not pay attention to text length, spelling, grammar and sentence structure but to genre elements when viewing or comparing the texts.

To initiate reflection, students were also asked which genre elements they could find in their first argumentative essay. Additionally, they were asked the following question: *"What would you change and/or add in/to your own essay to improve it?".*

4.4 Instruments

4.4.1 Text material for intervention

For the individual work of students (analyzing text exemplars), we selected 12 argumentative texts. These 12 texts utilized to generate exemplars for the intervention were deliberately chosen and carefully modified. In our previous intervention study (Mombaers et al., 2024), specific texts were adjusted to minimize distractions for students, such as spelling, style, structure, word choice, and so forth. Additionally, texts were manipulated to hold the desired ASRAW elements. The topic of the texts was broad and concerned whether or not to keep animals in zoos.

In the single exemplars condition, texts scoring above 65% on analytic assessment (with a mean score of 73.33%) and receiving holistic rankings between 9 and 95 out of 165 texts were chosen for analysis. For texts in the analogue selection, we selected those with total analytical scores of either 70% or 75% (with a mean score of 74.17%), with closely matched scores on different ASRAW aspects. Holistic rankings for texts in the analogue condition ranged from 1 to 85.

A comprehensive explanation of the selection and modification process of the text exemplars can be found in our previously published paper (Mombaers et al., 2024).

In selecting texts for teacher instruction, we chose three texts from another condition (near miss) as described in our previous study (Mombaers et al., 2024), a condition which we did not select in this study. For the analogue condition, we enhanced selected texts by adding signaling words, making style changes, and incorporating claim data (in one of the two texts). These modifications aimed to increase the ASRAW score and, as a result, to create suitable texts for making analogue comparisons. Similarly, we improved the text selected for instruction in the

single condition by adding signaling words, modifying style, incorporating claim data, and including a counterargument claim.

4.4.2 Motivation tests

Students have to read a significant number of texts during the instruction phase but also during the individual work. Therefore, one could argue that reading motivation might influence the enhancement of genre knowledge. Additionally, we also know from previous studies that writing motivation is a key predictor of text quality (e.g., De Smedt et al., 2016, 2023; García & de Caso, 2004; Graham et al., 2022; Troia et al., 2013). Hence, the Reading Motivation Questionnaire (RMQ) and the Writing Motivation Questionnaire (WMQ) were used to measure respectively reading and writing motivation. The WMQ is adapted from the Reading Motivation Questionnaire (Schiefele & Schaffner, 2016) that examines the same motives for reading through 34 items. For the WMQ, the number of items was reduced to 28 (Graham et al., 2022). Both questionnaires look at the motives for why students read or write (intrinsic or extrinsic motivation) and at the motives that drive students' reading or writing. It examines students' willingness to read or write because it is rewarding or satisfying in its own right (intrinsic motivation); makes it possible to achieve specific outcomes external to the activity or process of reading or writing writing (extrinsic motivation), and helps students regulate their emotions (self-regulatory motivation). Crohnbach's alpha for the reading (α = .93) and writing motivation tests (α = .95) indicated excellent internal consistency (Nunally & Bernstein, 1994).

The students were randomly assigned to the different conditions (i.e., based on their class group). Therefore, we did not expect differences in reading and writing motivation. Nevertheless, the motivation tests will be used to check whether students within the two conditions did not significantly differ in terms of reading and writing motivation.

4.4.3 Pretest

Pretesting included two aspects: a test for genre knowledge and writing a first argumentative essay.

An instrument comprising pretest components was developed to assess genre knowledge, which serves as one of the dependent variables in this study. Students were prompted in the pretest to offer advice to a friend on composing a well-crafted argumentative essay, to measure genre knowledge. Advising a novice friend in a particular writing genre has proven to be an effective method for evaluating students' writing proficiency (Schoonen & de Glopper, 1996), as demonstrated in various studies (e.g., Mombaers et al., 2024; Koster & Bouwer, 2016; van Drie et al., 2021). Students received this pretest in Qualtrics and were instructed as follows: *"Imagine your best friend needs to write an argumentative essay for school but has*

never done so before. Provide advice on the essential genre elements that should be included in their text to ensure it is effective. List as many genre elements (of an argumentative essay) as possible so that you can share this list with your friend."

Subsequently, students were asked to compose an argumentative essay regarding the topic of keeping animals in zoos, with the instruction:

"a) Write an argumentative essay about whether or not animals should be kept in zoos. Take a clear stance (for or against) and support your position with arguments. Your final text should be at least ¾ page long.
b) You will be provided with two source texts. You may select relevant information from these sources for inclusion in your own text. Additionally, you can incorporate your own ideas and perspectives."

4.4.4 Posttest

The posttest included two aspects: a test for genre knowledge (in Qualtrics) and writing a second argumentative essay. In the posttest for genre knowledge, students were asked whether they would revise the advice that they had given to their friend. If affirmative, they were then prompted to specify which genre elements they would add, modify, or remove from their list. Students also had to write another argumentative essay. This time, the topic was the right to vote from the age of 16. The instruction for this second argumentative essay was the same as the instruction for the first text (except for the topic). The students received the booklets in digital form (through the school's online platform) and had to fill them in digitally. The texts were given to them on paper. That way, they had the choice to examine them on their computer screens or on paper.

4.4.5 Retention test

Students were given a retention test 4 to 6 weeks after the intervention to test genre knowledge and text quality. The instructions for the genre knowledge retention test were exactly the same as the ones in the pretest and students completed this test in Qualtrics.

Students were instructed to write a third argumentative essay on the topic of stress at school. The instructions that students received for the retention test were identical to those given during the pretest. There were 66 missing essays in the data. This number is larger than for the first (26 missing) and second essay (36 missing) students wrote, primarily due to teachers not handing in this third essay.

Pre, post, and retention tests on genre knowledge were scored, considering different categories of genre knowledge. Table 2 shows the scoring weights for each aspect within this category. These categories and their scoring weights were developed in close collaboration with an expert in argumentative writing and proved to be an adequate way to assess genre knowledge in Mombaers et al. (2024).

Table 2. Categories of genre knowledge and their scores

Category	*Score in points*
ASRAW elements	3 or 4
reinforcement of argumentation	2
text goal	1
IME structure	2
general text structure	2
language use	1

The first author scored all students' pre and posttests on genre knowledge. The scores of each student for each category of genre knowledge was summed up to get a single score of genre knowledge for each student. In the posttest, students could add genre elements, make changes to the elements that they had already listed and/or delete genre elements. Adding more correct elements meant an increase in their posttest score compared to their pretest. Students' changes to genre elements could result in an increase or decrease in their score compared to their pretest. Deleting genre elements that were incorrect in their pretest did not affect their score, since they did not receive any points for these elements in the first place. But deleting correct genre elements, resulted in a decrease of their score.

The retention test held the same questions as the pretest, so this was scored in the same manner. Appendix B includes a more detailed list of categories of genre knowledge and their scores. The second author scored 10% of the pre, post and retention test. Interrater reliability was calculated through intraclass correlation (ICC). The ICC for the pretest was 0.97 (95%. CI: 0.92, 0.99), for the posttest it was 0.97 (95% CI: 0.93, 0.99), and for the retention test it was 0.96 (95% CI: 0.86, 0.99), indicating strong agreement among raters (Koo & Li, 2016).

Before the assessment, all texts were carefully 'cleaned' to correct spelling and punctuation errors, ensuring unbiased ratings from the evaluators.

In order to assess the argumentative essays written by students regarding text quality, three students, enrolled in university study programs that contained linguistics were enlisted to perform analytical scoring. These students underwent comprehensive training to ensure the reliability of their assessments. Initially, they were provided with background information about the study and the specific intervention. Following this, they were introduced to the ASRAW scoring instrument, which they would utilize to assess the essays. Their first task involved scoring five argumentative essays. Subsequently, they received detailed feedback on their scoring methodologies. Afterwards, they engaged in a second scoring task, assessing three additional texts. Once again, they received feedback to refine their scoring approaches. Following this iterative process, all participating students demonstrated proficiency in evaluating the essays according to the ASRAW criteria.

To assess inter-rater reliability, 10% of the texts were randomly selected for double scoring by one of the students who had not conducted the initial assessment. Using intraclass correlation (ICC), we evaluated the agreement between raters. The

resulting ICC was 0.73 (95% CI: 0.58, 0.83), suggesting a moderate level of agreement between the raters (Koo & Li, 2016). We considered this level of agreement acceptable, especially given the complexity of the skill being assessed.

4.5 Treatment fidelity

In this study, a lot of attention has been paid to the essential treatment fidelity aspects that a proper intervention has to contain (Capin, 2018; Dane & Schneider, 1998; Sanetti et al., 2021). Sanetti et al. (2021) highlight the importance of adherence, dosage, exposure and quality. Adherence refers to whether the intervention steps were implemented as planned. Dosage is the frequency with and duration for which the intervention is delivered, and exposure is the frequency with and duration for which a recipient received the intervention. Quality refers to how well the intervention steps were implemented. Also, participant responsiveness in the intervention is an important aspect of treatment fidelity (Capin, 2018; Dane & Schneider, 1998). Table 3 shows how these treatment fidelity guidelines were checked through open observations by the first author and the checklists and logbooks that teachers had to fill in.

Table 3. Treatment fidelity: data collection methods

Aspect of treatment fidelity	*Observation*	*Logbook & checklist*
adherence	adherence to the prescribed instruction (introduction - practice - reflection phase)	achievement of lesson objectives
dosage	duration of delivery of the intervention	all intervention steps implemented within the prescribed time indication?
exposure	the duration for which recipients received the intervention	the duration for which recipients received the intervention
quality	how well intervention steps were implemented:	clarity of the instructional materials
	interpersonal interactions	feasibility of the lessons
	sensitivity to students	suggested adjustments to the lessons
	effective class management	
participant responsiveness	the cooperation and attention of students during the intervention	involvement of students during the intervention

Checklists, logbooks and the observation instrument can be found on OSF (https://osf.io/3ht7e/?view_only=6da7d13f961046c99718d62463dee822).
Regarding the observations, we provide some more information in what follows.

The first author observed each teacher during at least one lesson period (lesson periods 3, 4, and/or 5) to ensure adherence to the study protocol. Additionally, the researcher assessed class management and student cooperation and involvement during these observations (see Table 3). However, there is one teacher that could not be observed due to communication issues. Instead, this teacher was interviewed online to discuss study protocol implementation in her lessons. During observations, if the researcher noted deviations from the lesson protocol, she prompted the teacher to realign with it (during or after the lesson). Examples of deviation included incomplete definitions or inadequate explanations of genre knowledge terms. Observation instruments were designed to facilitate an open observation approach, covering aspects such as time on task (individual and group), additional teacher support (both content-related and non-content related) and four thematic areas: implementation of intervention aspects, teacher instruction, class management, student involvement.

Based on the observations and the review of checklists and logbooks maintained by teachers, we can affirm that all teacher across all class groups met the treatment fidelity requirements, with the exception of one group. Data from the 9 students in this particular class group (Commercial Organization in school 3) were excluded from the analyses. Due to the intervention's poor quality, we could not assure that the requirements for adherence, dosage, and exposure were met. The class group presented some challenges for the teacher in terms of class management, resulting in less time on task for students during the intervention. Additionally, the classroom environment allowed for some disruptions, which made it difficult for all students to maintain their focus at all times.

We conducted multilevel linear mixed effects modelling in RStudio (R Core Team, 2021) on the data since students were nested in classrooms. So, classes was added to the models as a random effect. We decided not to include school as a random effect because only five schools were involved, making it hard to disentangle between classes and between school variances. Hence, we did not believe that the level school would make a difference in the results. All teachers received the same training and teacher manual to create uniformity within the intervention. In addition, observations did not show that teachers from different schools followed a different approach.

The packages tidyr (Wickham & Henry, 2019), tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), lmerTest (Kutznetsova et al., 2017), dplyr (Wickham et al., 2021), car (Fox & Weisberg, 2019), magrittr (Bache & Wickham, 2014), emmeans (Lenth, 2020), and MuMIn (Bartón, 2023) were used. The data that support the findings of this study and the analyses that were conducted, are openly available on OSF at https://osf.io/3ht7e/?view_only=6da7d13f961046c99718d62463dee822.

Descriptive statistics can be found in table 4 across conditions and table 5 (per condition).

Table 4. Descriptive statistics across conditions

variabele	*mean*	*SE*	*min*	*max*
genre_pre	7.48	5.83	0	21
genre_post	12.85	8.69	0	36
genre_ret	14.54	8.30	0	31
ASRAW_pre	4.44	4.04	0	18
ASRAW_post	7.54	6.14	0	22
ASRAW_ret	8.13	6.07	0	18
text1	57.13	25	0	95
text2	62.41	24.10	20	90
text3	40.65	25.85	0	95
reading_mot	63.54	15.50	38	105
writing_mot	46.85	15.67	28	87

Table 5. Descriptive statistics per condition

	SINGLE					*ANALOGUE*			
variabele	*mean*	*SD*	*min*	*max*	*variabele*	*mean*	*SD*	*min*	*max*
genre_pre	7.85	6.38	0	21	genre_pre	6.85	4.86	0	18
genre_post	14.41	8.84	0	36	genre_post	10.2	7.96	0	35
genre_ret	14.68	8.83	0	30	genre_ret	14.30	7.53	0	31
ASRAW_pre	5.00	4.64	0	18	ASRAW_pre	3.5	2.56	0	7
ASRAW_post	9.44	6.37	0	22	ASRAW_post	4.30	4.12	0	15
ASRAW_ret	8.71	6.51	0	18	ASRAW_pre	7.15	5.23	0	18
text1	60.59	22.86	20	95	text1	51.25	28.51	0	95
text2	63.68	25.56	20	90	text2	60.25	21.85	25	85
text3	40.00	28.71	0	95	text3	41.75	20.73	20	80
reading_mot	65.56	15.94	38	105	reading_mot	65.20	14.99	43	94
writing_mot	45.29	16.13	28	87	writing_mot	49.50	14.87	29	72

The values for reading and writing motivation are similar across the two conditions. This indicates that students' motivation in the single and analogue condition was not

that different, which is why we did not control for reading or writing motivation in our statistical models.

Multilevel linear mixed models were used to create models for (general) genre knowledge and knowledge on ASRAW elements on the one hand and text quality on the other hand as dependent variables. We chose to include models with ASRAW as a dependent variable to ascertain whether students in the single and analogue condition would learn more or less about the elements on which these exemplar texts were manipulated. Independent variables that were added to the models were time (3 measurement occasions), condition and the interaction between time and condition. For each dependent variable, we began by fitting a null model with only the class variable as a random effect. We then iteratively added different independent variables to this model to assess whether their inclusion improved the model (table 6).

For both genre knowledge and ASRAW we retained a model that only included time as an independent variable. These models all showed the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

For text quality, we selected the model that included time, condition and the interaction between time and condition. This model is statistically better than the model with only condition as a predictor ($p = <.05$) and the AIC is comparable to model 1. In addition, the effect size (marginal R^2) for the overall model 3 was slightly higher (0.14), than model 2 (0.12) and than model 1 (0.13), indicating that model 3 accounts for a slightly higher proportion of variance compared to the other two models. Finally, this model will be able to confirm or reject hypothesis 5: *Analyzing single or analogue text exemplars will not lead to differential improvements in text quality*.

For the complete structure of model building, we refer to the quarto file on OSF (https://osf.io/3ht7e/?view_only=6da7d13f961046c99718d62463dee822).

Table 6. Models fits for genre knowledge, ASRAW and text quality

		Genre knowledge					*Knowledge on ASRAW elements*					*Text quality*				
model	parameters	AIC	BIC	logLikelihood	df	p	AIC	BIC	logLikelihood	df	p	AIC	BIC	logLikelihood	df	p
1	5	985.24	1005.0	-487.62			1008.8	1028.6	-499.38			864.16	883.04	-427.08		
2	6	986.60	1010.4	-487.30	1	0.43	1010.8	1034.5	-499.38	1	0.99	866.10	888.75	-427.05	1	0.81
3	8	990.14	1021.8	-487.07	2	0.80	1013.3	1045.0	-498.66	2	0.48	864.06	894.26	-424.03	2	0.05*

5. RESULTS

5.1 Genre knowledge

To answer the first three hypotheses concerning genre knowledge, two linear mixed effects model for genre knowledge were generated. One was a model for general genre knowledge and the other one for knowledge on ASRAW elements (table 7).

Table 7. Mixed effects models for genre knowledge and for ASRAW elements

Genre knowledge					*ASRAW elements*		
Effect		*Estimate*	*SE*	*p value*	*Estimate*	*SE*	*p value*
Fixed effects							
	intercept	-0.54	0.14	0.00***	-0.50	0.14	0.000***
	time 2	0.81	0.10	0.00***	0.73	0.10	0.000***
	time 3	0.90	0.11	0.00***	0.74	0.11	0.000***
Random effects							
	class	0.21	0.46		0.21	0.45	

*Note 1. * p < 0.05, ** p < 0.01, *** p < 0.001.*

Note 2. Values shown in table are standardized.

5.1.1 General genre knowledge

The linear mixed model analysis investigated the impact of an intervention implemented between time points 1 (pretest), 2 (posttest) and 3 (retention test) on genre knowledge, while accounting for the nested structure of the data within classes (table 7).

Results revealed significant effects of the intervention since students demonstrated a significant increase in genre knowledge. Specifically, compared to time 1, genre knowledge increased substantially by an estimated 0.81 standard deviation at time 2 ($t(8.22)$, $p < .001$), and by an estimated 0.90 standard deviation at time 3 compared to time 1 ($t(8.50)$, $p < .001$).

In other words, students' genre knowledge significantly increased from pretest to posttest, and this progress was maintained at the retention test (0.90 sd). The effect sizes in Cohen's d at posttest (d=0.82) and retention test (d=0.85) were large (Cohen, 1988).

Drawing from the model outcomes, students' genre knowledge displayed a marginal increase of 0.09 standard deviations (from 0.81 to 0.90) between time 2 and time 3. However, this difference did not reach statistical significance according to the conducted post-hoc test (see table 13, Appendix C). Consequently, we can only derive conclusions regarding significant differences between the pretest and posttest, as well as between the pretest and retention test. Nonetheless, the findings

based on the outcomes of the selected model indicate that genre knowledge can indeed be enriched through learning from both single and analogue text exemplars. Thus, we can confidently confirm hypothesis 1 and 2: *Genre knowledge can be improved through learning from single (H1) and analogue (H2) text exemplars.*

Since the estimated model, which incorporates condition as a fixed effect both as a main effect and in interaction with time, did not exhibit significant improvement compared to the previously described model (see table 6), we can affirm that there are no discernible differential effects of the condition on genre knowledge across all time points. Whether analyzing single exemplars or analogue exemplars, there is no substantial difference in enhancing general genre knowledge. Hence, the specific condition students are in does not play a moderating role in predicting their genre knowledge.

5.1.2 ASRAW elements

The linear mixed model analysis investigated the impact of an intervention implemented between time points 1 (pretest), 2 (posttest) and 3 (retention test) on the knowledge of ASRAW elements, while accounting for the nested structure of the data within classes (table 7). Results revealed significant effects of the intervention since students demonstrated a significant increase in knowledge on ASRAW elements. Specifically, compared to time 1, ASRAW knowledge increased substantially by an estimated 0.73 standard deviation at time 2 ($t(7.18)$, $p < .001$), and by an estimated 0.74 standard deviation at time 3 compared to time 1 ($t(6.81)$, $p < .001$). In other words, students' knowledge on ASRAW elements significantly increased from pretest to posttest, and at the retention test students' knowledge on ASRAW elements was almost the same at posttest (0.73 sd compared to 0.74 sd). The effect sizes in Cohen's d at posttest (d=0.72) and retention test (d=0.68) were medium to large (Cohen, 1988). Post-hoc tests for time 2 and 3 did not show any significant differences between these two measurement occasions (see Appendix C, table 14).

Since the estimated model, which includes condition as a fixed effect both as a main effect and in interaction with time, did not exhibit significant improvement compared to the previously described model (table 6), we can affirm that there are no discernible differential effects of the condition on the knowledge of ASRAW elements across all time points. Whether analyzing single exemplars or analogue exemplars, there is no substantial difference in enhancing ASRAW knowledge. Hence, the specific condition students are in does not play a moderating role in predicting their knowledge on ASRAW elements.

Taking these results for general genre knowledge and knowledge on ASRAW elements into consideration, we can conclude that *analyzing single and analogue exemplars results in similar levels of improvement in general genre knowledge ánd ASRAW elements* (at both time 2 and 3), thereby confirming *hypothesis 3*.

5.2 Text quality

Another linear mixed effects model was selected to examine text quality as a dependent variable to confirm or reject hypotheses 4 and 5 (table 8).

Table 8. Mixed effects model for text quality

Effect		*Estimate*	*SE*	*p value*
Fixed effects				
	intercept	-0.16	0.17	0.34
	time 2	0.48	0.18	0.01**
	time 3	-0.38	0.21	0.07
	single	0.37	0.23	0.12
	Time2:single	-0.52	0.23	0.03*
	Time3:single	-0.52	0.27	0.05.
Random effects				
	class	0.09	0.30	

*Note 1. . p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.*

Note 2. Values shown in table are standardized.

Note 3. The analogue condition is the reference condition.

The selected linear mixed model (table 8) aimed to assess the influence of the intervention on text quality, conducted between time points 1 (pretest) and 2 (posttest), and to evaluate the retention of effects at time 3 (retention test). In addition, the model examined whether there were differential effects by including both condition and the interaction between time and condition in the analysis. The model accounted for the nested structure of the data within classes.

The results show that the intercept, representing the estimated mean for text quality for the analogue condition at the pretest was not statistically significant (t(-0.97) = -.10 , p = 0.34). Moreover, students in the single condition do not score significantly higher on text quality than students in the analogue condition on pretest (p=0.12).

At time 2 we found a significant positive effect on text quality (t(2.73), p = 0.01), indicating an improvement in text quality from pretest to posttest across conditions, with a small effect size (d=0.27). On average and at posttest, students showed an improved text quality, 0.48 sd higher than students on the pretest in the analogue condition (cf. intercept).

Nevertheless, we do not see a positive effect of time in all conditions, because for the single condition we do not see the same growth at time 2. So, students from the single condition do not necessarily show better text quality at time 2. This is shown by the interaction between time and condition, which indicated a significant negative effect for the single condition at time 2 (t(-2.23), p = 0.3), meaning that

students in the single condition scored approximately 0.04 sd lower (0 + 0.48 + 0 + 0 – 0.52= -0.04) than average on text quality at posttest, with a small effect size (d=-0.22). Students in the analogue condition, scored 0.48 sd higher than average (0 + 0.48 + 0 + 0 + 0= 0.48) on text quality at posttest.

The effect at the retention test (time 3) was negative and approached significance (t(-1.84), p = 0.07), suggesting a decline in text quality from pretest to retention test of -0.37 sd, although this decline was not statistically significant with a small effect size (d=-0.18).

The condition-specific effects were not significant on their own (single: t(1.58), = .16 , p = 0.12), indicating no significant difference in text quality at the pretest between the single and the analogue condition. For the retention test, the interaction term (single condition at time 3) approached significance (t(-0.52), p = 0.05), suggesting a potential trend towards a greater decline in text quality from pretest to retention test in the single condition compared to the analogue condition, with a small effect size (d=-0.20).

Important to note is that there was a high number of third texts from students missing from the data. For the retention test, we did not receive texts from 66 participating students. This was a higher number than for posttest (36 texts missing) and for pretest (26). The high number of missing values for the retention test might have influenced the analyses, leading to just a 'nearly significant' outcome at time 3.

In summary, the intervention significantly improved text quality with students in the analogue condition from pretest to posttest. Students in the single condition did not write better quality texts at posttest than compared to their pretest.

Based on these findings, the effectiveness of the intervention in enhancing text quality is evident, but only from pretest to posttest and for students in the analogue condition.

For the retention test, there appears to be a decline of text quality for both conditions, with a potentially greater decline in the single condition. However, these observations should be interpreted with caution as the estimates did not reach full statistical significance. Additionally, the relatively small effect sizes may be attributed to the missing data.

Therefore, hypothesis 4, which *posits that text quality can be improved through single and analogue text exemplars*, can only partially be confirmed, from pretest to posttest and for the analogue condition. The significant interaction effect between condition and time at posttest indicates differential improvements in text quality depending on whether single or analogue text exemplars were analyzed. Consequently, hypothesis 5, which anticipated no differential effect by condition, is not confirmed. Students' text quality in the analogue condition improved more than that of their peers in the single condition from pretest to posttest. At the retention test, students in the single condition seemed to experience a greater decline in text quality than those in the analogue condition, but this result should be viewed with caution due to the lack of full statistical significance.

5.3 Correlations

Pearson's correlations coefficients between general genre knowledge or knowledge on ASRAW elements and time (per condition) were calculated in order to either confirm or refute hypothesis 6. We chose to explore correlations both across conditions and within each condition. Given the observed variations in text quality between conditions, we anticipated differences in the correlations between genre knowledge and text quality as well.

5.3.1 Correlations for general genre knowledge

Table 9. Pearson's correlations between genre knowledge and text quality

time	*correlation coefficient*	*df*	*t-value*	*p-value*
1	0.04	113	0.44	0.66
2	0.16	112	1.73	0.09
3	0.23	76	2.06	0.04*

Note. n = 134

Table 9 presents Pearson's correlations between genre knowledge and text quality over three different time points (pretest, posttest and retention test) across conditions.

At pretest, the correlation coefficient of 0.04 indicates a very weak (de Vaus, 2002) and statistically non-significant correlation between genre knowledge and text quality at this initial time point. At posttest, the correlation coefficient is 0.16 and the p-value is 0.09. Although the correlation is still relatively weak, there is a trend towards significance, suggesting a potential increase in the relationship between genre knowledge and text quality. By the retention test, the correlation coefficient further increased to 0.23 with a p-value of 0.04. This correlation is statistically significant at the 0.05 level, suggesting a modest but meaningful positive relationship between genre knowledge and text quality at this time point.

These results suggest that the relationship between genre knowledge and text quality strengthens at the different time points. While the correlation is weak at pretest and posttest, by the retention test, there is a significant positive correlation, indicating that greater genre knowledge might be associated with higher text quality as participants progress through the study period.

Table 10. Pearson's correlation between genre knowledge and text quality per time & per condition

time	*condition*	*correlation coefficient*	*df*	*t-value*	*p-value*
1	single	0.15	64	1.22	0.23
1	analogue	-0.04	47	-0.29	0.78
2	single	0.13	63	1.09	0.28
2	analogue	0.19	47	1.35	0.18
3	single	0.23	45	1.59	0.12
3	analogue	0.23	29	1.28	0.21

Table 10 illustrates the Pearson's correlations between general genre knowledge and text quality at three different times and under two conditions: single and analogue. At time 1, the single condition shows a weak non-significant positive correlation, while the analogue condition shows a near-zero and non-significant correlation.

At time 2, the single condition again exhibits a weak positive correlation, and the analogue condition shows a lightly higher but still non-significant correlation.

At time 3, both conditions show a modest positive correlation, though these correlations are not statistically significant.

These results indicate that, while there are weak to modest positive correlations between general genre knowledge and text quality in both conditions over time, none of these correlations reach statistical significance within the conditions at any given time point.

5.3.2 Correlations for knowledge on ASRAW elements

Since we only discovered a significant positive correlation with general genre knowledge at time 3, we considered that a more detailed delineation of genre knowledge might yield additional insights into these correlations. Particularly, our assessment of text quality using ASRAW criteria did not encompass factors such as text structure, text goal, or the reinforcement of argumentation in the evaluation of students' argumentative essays. However, these elements are integral components of broader genre knowledge. Hence, we decided to narrow down our focus to specific genre knowledge of ASRAW elements, suspecting that they might exhibit stronger and more significant correlations.

Table 11. Pearson's correlation between ASRAW and text quality per time

time	*correlation coefficient*	*df*	*t-value*	*p-value*
1	0.11	113	1.14	0.25
2	0.16	112	1.71	0.09
3	0.24	76	2.16	0.03*

Table 11 presents Pearson's correlations between knowledge on ASRAW elements and text quality at the same three time points.

At time 1 and 2, the correlation coefficient indicates a weak and statistically non-significant correlation between knowledge on ASRAW elements and text quality at this initial time point. By time 3, the correlation coefficient increases to 0.24. This correlation is statistically significant at the 0.05 level, suggesting a modest, but meaningful positive relationship between knowledge on ASRAW elements and text quality at this later time point.

These results indicate that the relationship between the knowledge on ASRAW elements and text quality strengthens over time. While the correlation is weak at pretest and posttest, by the retention test, there is a significant positive correlation, indicating that greater knowledge of ASRAW elements is linked to higher text quality as students advance through the study period.

Table 12. Pearson's correlation between ASRAW and text quality per time & per condition

time	*condition*	*correlation coefficient*	*df*	*t-value*	*p-value*
1	single	0.18	64	1.45	0.15
1	analogue	0.03	47	0.20	0.84
2	single	0.06	63	0.51	0.61
2	analogue	0.31	47	2.22	0.03*
3	single	0.21	45	1.43	0.16
3	analogue	0.34	29	1.93	0.06

Table 12 shows Pearson's correlations between knowledge on ASRAW elements and text quality at three different times and under two conditions: single and analogue.

The single condition shows a weak positive correlation at time 1, while the analogue condition shows a near-zero correlation, both of which are statistically non-significant. At time 2, the single condition again exhibits a very weak positive correlation. However, the analogue condition shows a moderate positive correlation, which is statistically significant, indicating that greater knowledge on ASRAW elements is associated with higher text quality in the analogue condition at this time point. The single condition shows a modest positive correlation at time 3, while the analogue condition shows a stronger positive correlation, approaching statistical significance.

These results suggest that the relationship between knowledge on ASRAW and text quality varies by condition and over time, with the analogue condition showing a significant positive correlation at time 2 and a trend towards significance at time 3.

Overall, the data suggests that while initial correlations between genre knowledge and text quality are weak, there is a trend towards a stronger and significant positive correlations over time, particularly for the knowledge of ASRAW elements and within the analogue condition. Also, across conditions we see a significant correlation between time and general genre knowledge or knowledge on ASRAW elements at the retention test. For the single condition, the data does not show any significant correlations. As a result, we can only partially support the hypothesis that *genre knowledge is correlated with text quality*.

6. CONCLUSION & DISCUSSION

The first and most important conclusion of this paper is that improving text quality can be achieved by analyzing analogue text exemplars. Focusing solely on enhancing genre knowledge when analyzing analogue text exemplars, without attempting to improve any other influential factors, results in improved text quality from pretest to posttest. Learning from both single and analogue text exemplars enhances genre knowledge, and for the analogue exemplars, it also improves text quality.

There is no significant difference in the acquisition of genre knowledge, confirming the findings of our previous study (Mombaers et al., 2024), which applies to both general genre knowledge or the knowledge on specific ASRAW elements. However, different outcomes are observed in terms of text quality between students in the single and analogue condition. Specifically, students who studied analogue text exemplars produced higher quality argumentative essays compared to their peers in the single condition. This finding is intriguing, as one might expect that without differences in genre knowledge acquisition, the improvement in text quality would be similar. A possible explanation is that the transfer of knowledge occurs more smoothly when students analyze analogue text exemplars. We can thus confirm that learning from exemplars can improve complex skills (e.g., Alfieri et al., 2013; Gadgil & Nokes, 2009; Kurtz et al., 2001; Mombaers et al., 2024) such as writing. Particularly, in the knowledge transfer to a complex skill like writing, analogue exemplars are better suited than single exemplars. By comparing two texts, students likely pick up additional elements beyond genre knowledge, which can improve their writing. This is supported by the relatively low positive correlation found between genre knowledge and text quality at posttest for the analogue condition. This suggests that other factors also play an important role in text quality. Given that text quality was assessed based on criteria derived from essential genre knowledge (Stapleton & Wu, 2015), one would expect a higher correlation. However, the ASRAW criteria list also considers the relevance and persuasiveness of (counter)arguments and rebuttals next to their presence. A student may include all necessary genre elements in their argumentative essay, but without ensuring their relevance and persuasiveness, their score will not be high. After all, being convincing is the primary goal of an argumentative essay. As mentioned above, genre knowledge alone is not sufficient to enhance text quality; other factors, such as the relevance and persuasiveness of an argument, also play a key role in determining the quality of an argumentative essay. Moreover, the correlation in the study of Olinghouse et al. (2015) between genre knowledge and text quality was higher (0.51), but in that study, text quality was assessed holistically, considering genre elements, but also the development of ideas, organization, sentence structure and word choice. This feeds the assumption that next to genre elements, other aspects are important in text quality and should be examined further to be identified. Future studies should also examine the differences in the transfer of genre knowledge to text quality using measures such as think-aloud protocols or eye-tracking combined

with cued recall to reveal the learning process. This may give us insight in the transfer of genre knowledge to writing. Additionally, future research could explore differences between students in the single and analogue conditions regarding the persuasiveness of their essays and the specific scores they received to explain the differential outcomes in text quality.

Despite the clear improvements in text quality at posttest, the results for the retention test were inconclusive. Both groups of students wrote lower quality texts compared to pretest, with students in the single condition performing even worse than those in the analogue condition. This is surprising given the noticeable retention of genre knowledge in both conditions, which again, indicates that the relationship between genre knowledge and text quality does not seem very strong. Students possessed the necessary genre knowledge to write high-quality argumentative essays. One possible reason for the lack of improvement in text quality could be the content of the source texts provided. The pretest and posttest source texts contained both information and opinions on the topic, whereas the retention test source texts were slightly different. The first retention source text provided opinions on stress at school, while the second primarily provided information backed by numerical data. This led students to incorporate data into their texts, and while doing so, neglecting the need for sound arguments, therefore resulting in texts that were informative rather than persuasive. Another reason for the decline in text quality could be students' motivation and the fact that they are not familiar with writing several texts within the same genre. Writing three texts in a relatively short period of time (from the students' perspective) seemed to overwhelm several students, as noted by several teachers in their logbooks. Students in Flanders are not used to writing several texts in the same genre within a relative short period of time. However, writing at least three texts in a specific genre to assess text quality in this genre is a guideline for generalizability in genre-specific writing (Bouwer & Van den Bergh, 2015; Kim et al., 2017; Schoonen, 2005, 2012; Van den Bergh et al., 2012) and is standard practice within writing research (e.g., Landrieu et al., 2024; Vandermeulen et al., 2023). In future studies employing text exemplars to enhance text quality, researchers should carefully consider the source texts used. Each should contain both information and opinions on the topic. Additionally, enhancing students' genre knowledge alone may not be sufficient to maintain text quality over time; motivation also seems to be crucial. This is backed by several writing researchers that highlight the importance of motivation (e.g., De Smedt et al., 2018; Graham et al., 2022; Zimmerman & Risemberg, 1997). Therefore, future intervention research set up to enhance text quality should include strategies to keep students motivated throughout the study.

7. LIMITATIONS

Despite its relevance to the field of learning from exemplars and argumentative writing, this study also contains some limitations. The first limitation is that through

missing data at retention test, we had to perform statistical analyses on a smaller dataset. This may have limited our generalizations for the retention of text quality and this leads to cautiousness in interpreting these results. However, all efforts were made to receive missing texts from the participating teachers, in which our efforts were successful in retaining some missing texts from post-test, but we were unable to get more retention texts.

A second limitation lies in the fact that we only let students write in one genre, namely the argumentative essay, so we cannot generalize these results on text quality for other text genres. As a last limitation, we need to acknowledge that we did not have a control group in this study. We chose not to include a control group since sequential conditions are favorable to control groups because they differ from the comparison condition only in that cases are studied in succession (Rittle-Johnson & Star, 2007). Nevertheless, the lack of a control condition should be considered when interpreting the results of this study.

The third limitation includes the lack of observation of the lessons of one teacher due to communication issues. While the absence of direct observation - and thus reliance on the teacher's self-report - may have limited the ability to objectively assess all aspects of protocol adherence, the detailed information provided during the interview offered valuable insights into the teacher's practices.

8. CONTRIBUTIONS

Though this study holds several limitations, it makes a substantial contribution to the domain of learning from (comparing) exemplars and argumentative writing research. The research contributes significantly to theory by demonstrating that learning from text exemplars can not only enhance genre knowledge but also leads to the production of higher quality argumentative texts. This finding is important as it validates the use of exemplars an effective educational tool, showing that when students are exposed to well-crafted examples of argumentative writing, they seem to be able to internalize the structural persuasive features that define the genre. Furthermore, the study underscores that a focused approach to improving one aspect of writing, namely genre knowledge, can have a broader impact on text quality. By concentrating on genre-specific elements when analyzing text exemplars, learners are better equipped to produce coherent and persuasive argumentative essays. This targeted learning approach does not only increase students' understanding of the genre but also enhances their ability to apply this knowledge in practice, resulting in higher quality argumentative essays, especially when genre knowledge is improved through analogue comparisons. This research provides a robust foundation for future studies aiming to explore the benefits of exemplar-based learning and its potential to elevate learners' writing standards.

9. IMPLICATIONS FOR PRACTICE

The findings of this study have several important implications for educational practice. First, if educators want to enhance genre knowledge, they can use either single or analogue exemplars to do so. However, when the goal is to improve students' argumentative essay writing, they should employ analogue text exemplars. This approach indicates that enhancing genre knowledge alone, without focusing on other writing aspects, can still lead to better argumentative essays when analogue text exemplars are used. Another implication is that source texts should be chosen carefully. In order to be used as an inspiration for an argumentative essay, each source text should contain both information and opinions on the topic. Moreover, while genre knowledge is crucial, it is equally important to address the relevance and persuasiveness of arguments. Students should be taught how to build a relevant and persuasive argument; this will improve the overall persuasiveness of their argumentative essays. Lastly, maintaining student motivation throughout the learning process is essential for sustained improvement in text quality. Educational strategies should, therefore, include motivational components, such as varied and engaging writing prompts, to keep students invested in their writing tasks over time.

In essence, this study underscores the multifaceted nature of effective writing instruction, highlighting the importance of integrating genre knowledge with attention to argumentative structure, source text selection, and sustained student motivation. By employing a holistic approach that addresses these elements, educators can foster substantial improvements in students' argumentative essay quality.

AUTHORS' NOTE

This work was supported by the FWO (Fonds voor Wetenschappelijk Onderzoek Vlaanderen – Flemish Research Fund) under Grant 1SE4823N.

REFERENCES

Abbuhl, R. (2011). Using models in writing instruction: A comparison with native and nonnative speakers of English. *SAGE Open, 1*(3), 1–12. https://doi.org/10.1177/2158244011426295.

Abdollahzadeh, E., Farsani, M. A., & Beikmohammadi, M. (2017). Argumentative writing behavior of graduate EFL learners, *Argumentation, 31*(4), 641-661. https://doi.org/10.1007/s10503-016-9415-5

Abrams, Z. (2019). Collaborative writing and text quality in Google docs. *Language Learning & Technology, 23*(2), 22–42. https://doi.org/10125/44681

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

Allagui, B. (2021). Writing a descriptive paragraph using an augmented reality application: An evaluation of students' performance and attitudes. *Tech Know Learn, 26*, 687-710. https://doi.org/10.1007/s10758-019-09429-2

Bacha, N. N. (2010). Teaching the academic argument in a university EFL environment. *Journal of English for Academic Purposes, 9*, 229-241. https://doi.org/10.1016/j.jeap.2010.05.001

Bache, S. M., & Wickham, H. (2014). Magrittr: A forward-pipe operator for R. R package version 2.0.1. Retrieved from https://CRAN.R-project.org/package=magrittr

Bañales, G., Ahumada, S., Martínez, R., Martínez, M., & Messina, P. (2018). Investigaciones de la escritura en la educación básica en Chile: Revisión de una década [Examination of writing in elementary school in Chile: Revision of one decade] (2007-2016). RLA. *Revista de Lingüística Teórica y Aplicada, 56*(1), 59-84. https://doi.org/10.4067/S0718- 48832018000100059

Bartoń, K. (2023). MuMIn: Multi-Model Inference. R package version 1.47.1. https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Beers, S., & Nagy, W. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing: An Interdisciplinary Journal, 22*, 185- 200. https://doi.org/10.1007/s11145-007-9107-5

Beaufort, A. (2007). *College writing and beyond: A new framework for university writing instruction*: Utah State University Press. https://doi.org/10.2307/j.ctt4cgnk0

Beauvais, C., Olive, T., & Passerault, J. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology, 103*(2), 415-428. https://doi.org/10.1037/a0022545

Bigger, A. T. (2022). Secondary students' proficiency with and knowledge of process and genre approaches to writing [Unpublished doctoral dissertation]. University of Pittsburgh.

Bouwer, R., & Van den Bergh, H. (2015). Toetsen van schrijfvaardigheid: hoeveel beoordelaars, hoeveel taken? [Testing writing skills: How many raters, how many tasks?]. *Levende Talen Tijdschrift*, *16*(3), 3–12. Retrieved from https://lt-tijdschriften.nl/ojs/index.php/ltt/article/view/1516

Capin, P., Walker, M. A., Vaughn, S., & Wanzek, J. (2018). Examining how treatment fidelity is supported, measured, and reported in K–3 reading intervention research. *Educational Psychology Review, 30*, 885-919. https://doi.org/10.1007/s10648-017-9429-z

Chambliss, M. J., & Murphy, P. K. (2010). Fourth and fifth graders representing the argument structure in written texts, *Discourse Processes, 34*(1), 91-115. https://doi.org/10.1207/S15326950DP3401_4

Carless, D., & Chan, K. K. H. (2016). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education*, *42*(6), 930–941. https://doi.org/10.1080/02602938.2016.1211246

Charney, D., & Carlson, R. (1995). Learning to write in a genre: What student writers take from model texts. *Research in the Teaching of English, 29*(1), 88–125. https://doi.org/10.58680/rte199515358

Childers, J. B. (2008). The structural alignment and comparison of events in verb acquisition *Proceedings of the Annual Meeting of the Cognitive Science Society*, 30. Retrieved from https://escholarship.org/uc/item/2jz052xr

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research, 7*(3), 351-370. https://doi.org/10.17239/jowr-2016.07.3.02

da Cunha, I., & Montané, M. (2019). Textual genres and writing difficulties in specialized domains. *Revista Signos: estudios de lingüística, 52*(99), 4-30. https://doi.org/10.4067/S0718-09342019000100004

De Smedt, F., Graham, S., & Van Keer, H. (2018). The bright and dark side of writing motivation: Effects of explicit instruction and peer assistance. *Journal of Educational Research, 112*, 152-167. https://doi.org/10.1080/00220671.2018.1461598

De Vaus, D. (2002). *Analysing social science data.* Sage Publications Limited.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*. https://doi.org/10.1016/S0272-7358(97)00043-3

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher, and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing, 29*, 833-868. https://doi.org/10.1007/s11145-015-9590-z

De Smedt, F., & Van Keer, H. (2018). Fostering writing in upper primary grades: A study into the distinct and combined impact of explicit instruction and peer assistance. *Reading and Writing*, *31*(2), 325–354. https://doi.org/10.1007/s11145-017-9787-4

De Smedt, F., Landrieu, Y., De Wever, B., & Van Keer, H. (2023) The role of writing motives in the interplay between implicit theories, achievement goals, self-efficacy, and writing performance. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1149923

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal Technology, Learning and Assessment, 5*(1), 1-35.

Donovan, C. A., & Smolkin, L. B. (2006). Children's understanding of genre and writing development. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing researc*h, pp. 131–143. The Guilford Press.

Faul, F., Erdfelder, E., Lang, AG. *et al.* G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39, 175–191 (2007). https://doi.org/10.3758/BF03193146

Ferretti, R. P., & De La Paz, S. (2011). On the comprehension and production of written texts: Instructional activities that support content-area literacy. In R. O'Connor & P. Vadasy (Eds.), *Handbook of reading interventions* (pp. 326–355). Guilford.

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology, 101*(3), 577–589. https://doi.org/10.1037/a0014702

Ferretti, R. P., and Lewis, W. E. (2013). Best practices in teaching argumentative writing, in S. Graham, C. A. MacArthur, and J. Fitzgerald (Eds.), *Best practices in writing instruction* (2nd ed, pp. 113-140). Guilford Press.

Ferretti, R. P., & Lewis, W. E. (2019). Best practices in teaching argumentative writing. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (3rd ed., pp. 135–161). Guilford Press.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Freedman, A., & Pringle, I. (1984). Why students can't write arguments, *English in Education, 18*(2), 73-84. https://doi.org/10.1111/j.1754-8845.1984.tb00668.x

Gadgil, S., & Nokes, T. (2009). Analogical scaffolding in collaborative learning. Proceedings of the annual Meeting of the Cognitive Science Society, 31. Retrieved from https://escholarship.org/uc/item/5b74x3jv

García, J.-N., & de Caso, A. M. (2004). Effects of a motivational intervention for improving the writing of children with learning disabilities. *Learning Disability Quarterly, 27*(3), 141-159. https://doi.org/10.2307/1593665

Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, D., & Namy L. L. (1999). Comparison in the development of categories. *Cognitive Development, 14*(4), 487-513). https://doi.org/10.1016/S0885-2014(99)00016-7

Gillespie, A., Olinghouse, N. G., & Graham, S. (2013). Fifth grade students' knowledge about writing process and writing genres. *The Elementary School Journal, 113*(4), 565-588. https://doi.org/10.1086/669938

Graham, S., Harbaugh-Schattenkirk, A. G., Aitken, A., Harris, K. R., Ng, C., Ray, A., Wilson, J. M., & Wdowin, J. (2022). Writing motivation questionnaire: Validation and application as a formative assessment. *Assessment in Education: Principles, Policy and Practice*, *29*(2), 238-261. https://doi.org/10.1080/0969594X.2022.2080178

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(4), 879-896. https://psycnet.apa.org/doi/10.1037/a0029185

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools.* Carnegie Corporation.

Granado-Peinado, M., Mateos, M., Martín, E., & Cuevas, I. (2019). Teaching to write collaborative argumentative syntheses in higher education. *Reading and Writing*, *32*(8), 2037–2058.

https://doi.org/10.1007/s11145-019-09939-6

Guo, L., Crossley, S., & McNamara, D. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Writing Assessment, 18*, 218-238. https://doi.org/10.1016/j.asw.2013.05.002

Hammer, R., Bar-Hillel, A., Hertz, T., Weinshall, D., & Hochstein, S. (2008). Comparison processes in category learning: From theory to behavior. *Brain Research*, *1225*, 102–118. https://doi.org/10.1016/j.brainres.2008.04.079

Hendry, G. D., Armstrong, S., & Bromberger, N. (2011). Implementing standards-based assessment effectively, incorporating discussion of exemplars into classroom teaching. *Assessment & Evaluation in Higher Education, 37*(2), 149-161. https://doi.org/10.1080.02602938.2010.515014

Hox, J., Moerbeek, M, & van de Schoot, R. (2017). *Multilevel Analyses, Tecniques and Applications. Quantative Methodology Series (3rd ed.).* Athenaeum Uitgeverij. https://doi.org/10.4324/9781315650982

Hyland, K. (2007). *Genre and second language writing*. University of Michigan Press.

Hyon, S. (2001). Long-term effects of genre-based instruction: a follow-up study of an EAP reading course, *English for Specific Purposes, 20*(1), 417-438. https://doi.org/10.1016/S0889-4906(01)00019-9

Hyon, S. (2002). Genre and ESL reading: A classroom study. In John, A. M. (Ed.), *Genre in the classroom*. Lawrence Earlbaum Associates Publishers.

Johnson, A., Wilson, J., & Roscoe, R. (2017). College student perceptions of writing errors, text quality, and author characteristics, *Assessing Writing, 34*, 72-84. https://doi.org/10.1016/j.asw.2017.10.002

Kim, Y.-S.G., Schatschneider, C., Wanzek, J., Gatlin, B., & Otaiba, S. A. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in Graders 3 and 4. *Reading and Writing, 30*(6), 1287-1310. https://doi.org/10.1007/s11145-017-9724-6

Koo, T. K, & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Koster, M., Tribushinina, E., De Jong, P.F., & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research, *Journal of Writing Research, 7*(2), 249-274. https://doi.org/10.17239/jowr-2015.07.02.2

Koster, M. P., & Bouwer, I. R. (2016). *Bringing writing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students* [Unpublished doctoral thesis]. Universiteit Utrecht. Retrieved from: https://dspace.library.uu.nl/bitstream/handle/1874/338041/BouwerKoster.pdf?sequence=1

Kurtz, K. J., Miao, C., & Gentner, D. (2001-). Learning by analogical Bootstrapping. *The Journal of the Learning Sciences, 10(*4), 417-446. https://doi.org/10.1207/S15327809JLS1004new_2

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Kuhn, D. (1999). A development model of critical thinking, *Educational Research, 28*, 16-46. https://doi.org/10.3102/0013189X028002016

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science, 22*(4), 545–552. https://doi.org/10.1177/0956797611402512

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Landrieu, Y., De Smedt, F., Van Keer, H. (2024). Argumentation in collaboration: the impact of explicit instruction and collaborative writing on secondary school students' argumentative writing. *Reading and Writing, 37*, 1407–1434 (2024). https://doi.org/10.1007/s11145-023-10439-x

Latifi, S., Noroozi, O., & Talaee, E. (2021). Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes, *British Journal of Educational Technology, 52*(2), 768-784. https://doi.org/10.1111/bjet.13054

Lee, J. J., & Deakin, L. (2016) Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays, *Journal of Second Language Writing, 33*, 21-34, https://doi.org/10.1016/j.jslw.2016.06.004.

Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means. R package version 1.5.4. Retrieved from https://CRAN.R-project.org/package=emmeans

Lipnevich, A., McCallen, L., Miles, K., et al. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559. https://doi.org/10.1007/s11251-013-9299-9

Martin, J. R. (2009). Genre and language learning: A social semiotic perspective, *Linguistics and Education, 20*(1), 10-21. https://doi.org/10.1016/j.linged.2009.01.003

McCann, T. M. (1989). Student argumentative writing knowledge and ability at the three grade levels, *Research in the Teaching of English, 23*(1), 62-76. https://doi.org/10.58680/rte198915528

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. Journal of Memory and Language, 25(4). 431-444. https://doi.org/10.1016/0749-596X(86)90036-7

Miller, J. G. (1984). Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology, 46*(5), 961-978. https://doi.org/10.1037/0022-3514.46.5.961

Miller, C. (2009, March). *What is a new genre?* [Paper presentation]. The College Composition and Communication, San Francisco, CA.

Mohsen, M. A., & Qassem, M. (2020). Analyses of L2 learners' text writing strategy: Process-oriented perspective, *Journal of Psycholinguistic Research, 49*(1), 435-451. https://doi.org/10.1007/s10936-020-09693-9

Mombaers, T., Van Gasse, R., & De Maeyer, S. (2024). Learning from comPA(I)Ring exemplars: Enhancing genre knowledge of argumentative texts. *Journal of Writing Research, 16*(1),163-193. https://doi.org/10.17239/jowr-2024.16.01.06

Moschella, J. A. (2023). *Computer-based formative assessment in developmental writing targeting argumentative essay writing skills*, [Doctoral dissertation]. Graduate School of Education Rutgers, The State University of New Jersey. ProQuest Dissertations & Theses Global.

Muller Mirza, N., & Perret-Clermont, A.-N. (2009) (Eds.). *Argumentation and Education*. Springer.

Namy, L. L., Gentner, D., & Clepper, L. E. (2007). How close is too close? Alignment and perceptual similarity in children's categorization. *Cognition, Brain, Behavior, 11*, 647–659.

NCES (2012). *The nation's report card: Writing 2011*. NCES.

Newell, G. E., Beach, R., Smith, J., & Vanderheide, J. (2011). Review of research: Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly*, *46*(3), 273–304. https://doi.org/10.1598/RRQ.46.3.4

Nunnally, J. C., & Bernstein, I. C. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, *97*(2), 157–169. https://doi.org/10.1037/0022-0663.97.2.157

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *Journal of Experimental Education*, *76*(1), 59–92. https://doi.org/10.3200/JEXE.76.1.59-92

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*(1), 37-50.

Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107*(2), 391-406. https://psycnet.apa.org/doi/10.1037/a0037549

Olive, T., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisation, *Learning and Instruction, 19*(4), 299-308, https://doi.org/10.1016/j.learninstruc.2008.05.005.

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. Voss, D. Perkins & J. Segal (Eds.), *Informal reasoning and education* (pp. 83–105). Erlbaum.

Pessoa, S., Mitchell, T. D., Miller, R. T. (2017), Emergent arguments: A functional approach to analyzing student challenges with the argument genre, *Journal of Second Language Writing, 38,* 42-55, https://doi.org/10.1016/j.jslw.2017.10.013.

Prata, M. J., de Sousa, B., Festas, I., & Oliveira, A. L. (2019). Cooperative methods and self-regulated strategies development for argumentative writing. *Journal of Educational Research*, *112*(1), 12–27. https://doi.org/10.1080/00220671.2018.1427037

Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, *38*(3), 444–456. https://doi.org/10.1016/j.system.2010.06.012

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Retrieved from https://www.R-project.org/.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561–574. https://doi.org/10.1037/0022-0663.99.3.561

Reed, S.K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 124-139. https://doi.org/10.1037/0278-7393.13.1.124

Ross, B. H., & Kennedy, P.T. (1990). Generalizing from the use of earlier examples in problem solving. Journal of Experimental Psychology: *Learning, Memory, and Cognition, 16*, 42-55. https://doi.org/10.1037/0278-7393.16.1.42

Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading & Writing Quarterly, 23*(3), 231-247. https://doi.org/10.1080/10573560701277575

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 93*(3), 447-472. https://doi.org/10.1002/sce.20276

Sanetti, L. M. H., Cook, B. G., & Cook, L. (2021). Treatment fidelity: What it is and why it matters. *Learning Disabilities Research & Practice, 36*(1), 5-11. https://doi.org/10.1111/ldrp.12238

Schiefele, U., & Schaffner, E. (2016). Factorial and construct validity of a new instrument for the assessment of reading motivation. *Reading Research Quarterly, 51*(2), 221–237. https://doi.org/10.1002/rrq.134

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

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam & H. Van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 1-22). Brill. https://doi.org/10.1108/S1572-6304(2012)0000027005

Schoonen, R., & de Glopper, K., (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. van den Berg & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 87-107). Amsterdam University Press.

Simon, S. (2008). Using Toulmin's Argument Pattern in the evaluation of argumentation in school science. *International Journal of Research & Method in Education, 31*(3), 277-287. https://doi.org/10.1080/17437270802417176

Smith, L. A., & Gentner, D. (2014). *UC Merced proceedings of the annual meeting of the Cognitive Science: The role of difference - Detection in learning contrastive categories*.

Smyth, P., & Carless, D. (2020). Theorising how teachers manage the use of exemplars: towards mediated learning from exemplars. *Assessment & Evaluation in Higher Education, 46*(3), 393-406. https://doi.org/10.1080/02602938.2020.1781785

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

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written Communication, 18*(4), 506–548. https://doi.org/10.1177/0741088301018004004

Stapleton, P., & Wu, Y. (A.). (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, *17*, 12–23. https://doi.org/10.1016/j.jeap.2014.11.006

Tardy, C. M. (2009). *Building genre knowledge.* Parlor Press.

To, J., Panadero, E., & Carless, D. (2022) A systematic review of the educational uses and effects of exemplars, *Assessment & Evaluation in Higher Education, 47*(8), 1167-1182. https://doi.org/10.1080/02602938.2021.2011134

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

Toulmin, S. (2003). *The uses of argument* (2nd ed.). Cambridge University Press.

Tribble, C. (2015). Writing academic English further along the road. What is happening now in EAP writing instruction?, *ELT Journal, 69*(4), 442-462. https://doi.org/10.1093/elt/ccv044

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2023). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading & Writing, 26*, 17–44. https://doi.org/10.1007/s11145-012-9379-2

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 Inisights into Theory, Methodology and Practices* (pp. 23-32). Brill. https://doi.org/10.1108/S1572-6304(2012)0000027005

Vandermeulen, N., Van Steendam, E., De Maeyer, S., Lesterhuis, M., & Rijlaarsdam, G. (2023). Learning to write syntheses: The effect of process feedback and of observing models on performance and process behaviors. *Reading and Writing, (37)*, 1375-1405. https://doi.org/10.1007/S11145-023-10483-7

Van Drie, J., Van Driel, J., & Van Weijen, D. (2021). Developing students' writing in History: Effects of a teacher-designed domain-specific writing instruction. *Journal of Writing Research,13*(2), 201-229. https://doi.org/10.17239/jowr-2021.13.02.01

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinsin, D., Seidel, D. P., Spinu, V.,… Yutani., H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A grammar of data manipulation. R package version 1.0.7. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). tidyr: Tidy Messy Data. R package version 1.0.0. Retrieved from https://CRAN.R-project.org/package=tidyr

Yasuda, S. (2011). Genre-based tasks in foreign language writing: Developing writers' genre awareness, linguistic knowledge, and writing competence. *Journal of Second Language Writing, 20*(2), 111–133. https://doi.org/10.1016/j.jslw.2011.03.001

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A Meta-Analysis, *Journal of Educational Computing Research, 61*(4), 723-924. https://doi.org/10.1177/07356331221127300

Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology, 22*(1), 73–101. https://doi.org/10.1006/ceps.1997.0919

APPENDIX A

Analytical scoring rubric for argumentative writing (Stapleton & Wu, 2015)

	Score: 5			Score: 0	
1. Claim(s) (5%)	States point(s) of view			Doesn't state point(s) of view	
	Score: 25	Score: 20	Score: 15	Score: 10	Score: 0
2. Data (25%)	a. Provides multiple reasons for the claim(s), and	a. Provides multiple reasons for the claim(s), and	a. Provides one to two reasons for the claim(s), and	a. Provides only one reason for the claim(s), or	a. No reasons are provided for the claim(s); or
	b. All reasons are sound/acceptable and free of irrelevancies.	b. Most reasons are sound/acceptable and free of irrelevancies, but one or two are weak.	b. Some reasons are sound/acceptable, but some are weak or irrelevant.	b. The reason provided are weak or irrelevant.	b. None of the reasons are relevant to/support the claim(s).
3. Counterargument Claim(s) / Alternative Point(s) of View (10%)	Score: 10			Score: 0	
	Provides counterargument(s) / alternative view(s)			Doesn't provide counterargument claim(s) / alternative view(s)	
	Score: 25	Score: 20	Score: 15	Score: 10	Score: 0
4. Counterargument Data / Supporting Reasons for Alternative Point(s) of View (25%)	a. Provides multiple reasons for the counterargument claim(s)/alternative view(s) and,	a. Provides multiple reasons for the counterargument claim(s)/alternative view(s), and	a. Provides one to two reasons for the counterargument claim(s)/alternative view(s), and	a. Provides only one reason for the counterargument claim(s)/alternative view(s), or	a. No reasons are provided for the counterargument claim(s)/alternative view(s), or

	b. All counterargument/reasons for the alternative view(s) are sound/acceptable and free of irrelevancies.	b. Most counterarguments/reasons for the alternative view(s) are sound/acceptable and free of irrelevancies, but one or two are weak.	b. Sound counterarguments/reasons for the alternative views(s) are sound/acceptable, mbut some are weak or irrelevant.	b. The counterargument/reason for the alternative view is weak or irrelevant.	b. None of the reasons are relevant to/support the counterargument claim(s)/alternative view(s).
5. Rebuttal Claim(s) (10%)	Score: 10			Score: 0	
	Provides rebuttal claim(s).			Doesn't provide rebuttal claim(s).	
	a. Refutes/points out the weaknesses of all the counterarguments, and	a. Refutes/points out the weakness of all the counterarguments, and	a. Refutes/points out the weaknesses of all the counterarguments, and	a. Refutes/points out the weaknesses of some counterarguments, or	a. No rebuttals are provided; or
	b. All rebuttals are sound/acceptable.	b. Most rebuttals are sound/acceptable, but one or two are weak.	b. Some rebuttals are sound/acceptable, but some are weak.	b. Few of the rebuttals are sound/acceptable; most of them are weak, or	b. None of the rebuttals can refute the counterarguments.
	c. The reasoning quality of all the rebuttals are stronger than that of the counterarguments.	c. The reasoning quality of most rebuttals are stronger than that of the counterarguments, while one or two are equal to that of the counterarguments.	c. The reasoning quality of some rebuttals are stronger than that of the counterarguments, while some are weaker than that of the counterarguments.	c. The reasoning quality of most rebuttals are weaker than that of the counterarguments.	

*Note: * An implicit requirement of rebuttal data is subsumed under the requirements of row 4 "Counterargument Data," that is, each piece of rebuttal data should be aligned with each piece of counterargument data in terms of both quantity and logic.*

APPENDIX B

Categories of genre knowledge and their score

1) *ASRAW* — *3 points*
   - position (for or against) (claim)
   - argumentation (claim data)
     - clear arguments
     - different arguments
     - arguments to support position
   - counterargument (counterargument claim) — *4 points*
   - argumenting counterargument (counterargument data)
   - refuting counterargument (rebuttal claim)
   - arguments for rebuttal (rebuttal data)
2) *Reinforcement of argumentation* — *2 points*
   - fact/opinion
   - describing pros and cons
   - giving examples
   - mentioning sources
   - providing proof
   - critical
   - objective/subjective
3) *Text goal* — *1 point*
   - persuading
4) *IME structure* — *2 points*
   - introduction
   - middle (body of the text)
   - ending / conclusion
5) *General text structure* — *2 points*
   - paragraphs
   - use of signal words
   - clearly structured
   - (good, attractive) title
   - Blank lines between paragraphs
6) *Language use* — *1 point*
   - formal language
   - proper Dutch (no dialect)
   - written in first person (I)
   - being aware of the audience that you are writing to
   - full sentences
   - clear & short sentences

APPENDIX C

Table 13. Overview of outcomes for post-hoc analyses on the difference between posttest and retention test for genre knowledge

time	*difference*	*SE*	*p-value*
3	-0.09	0.11	0.39

Table 14. Overview of outcomes for post-hoc analyses on the difference between posttest and retention test for knowledge on ASRAW elements

time	*difference*	*SE*	*p-value*
3	-0.02	0.11	0.88