Janssen, J., Erkens, G., & Kirschner, P.A. (2010). Effects of representational guidance during computer-supported collaborative learning. *Instructional Science*, *38*, 59-88.

EFFECTS OF REPRESENTATIONAL GUIDANCE DURING COMPUTER-SUPPORTED COLLABORATIVE LEARNING

Jeroen Janssen[*], Gijsbert Erkens[*], Paul Kirschner[†]

[*] Research Centre Learning in Interaction, Utrecht University, The Netherlands

[†] Centre for Learning Sciences and Technologies, Open University, The Netherlands

Janssen, J., Erkens, G., & Kirschner, P.A. (2010). Effects of representational guidance during computer-supported collaborative learning. *Instructional Science*, *38*, 59-88.

EFFECTS OF REPRESENTATIONAL GUIDANCE DURING COMPUTER-SUPPORTED COLLABORATIVE LEARNING

External representations (ERs) are graphical or diagrammatic representations that represent information about a topic (Van Bruggen, Kirschner, & Jochems, 2002; Zhang, 1997). A growing number of researchers advocates not only the use, but also the construction of ERs by students. The active construction of ERs, is thought to stimulate deeper and more mindful cognitive processing, resulting in larger learning gains (Ainsworth, 2006; Cox, 1999; Schnotz & Kürschner, 2008; Stern, Aprea, & Ebner, 2003). These ERs can be constructed individually or by groups of students. Co-construction of ERs may link the benefits of collaborative learning, such as fostering peer support, and critical discussion, to the benefits of constructing ERs (Munneke, Andriessen, Kanselaar, & Kirschner, 2007; Schwarz, Neuman, Gil, & Ilya, 2003; Toth, Suthers, & Lesgold, 2002). Unfortunately, this has not always been shown to be the case. Some researchers have reported unexpected results when students collaboratively construct ERs. Van Drie, Van Boxtel, Jaspers, and Kanselaar (2005) for example, attempted to stimulate historical reasoning and argumentation using Diagram, List, and Matrix representations, but found that students in their control condition (i.e., without co-construction of an ER) used more historical reasoning during their discussions.

Representational guidance refers to the fact that different ERs are capable of expressing different information (e.g., cause and effect relationships versus temporal relationships), make different information salient, or stimulate different cognitive processes than other ERs (Suthers & Hundhausen, 2003) and may explain the benefits (or lack thereof) of co-constructing ERs. Several studies investigating the effects of different argumentation tools showed representational guidance can influence students' behavior and learning process (Schwarz et al., 2003; Suthers, 2001; Van Amelsvoort, Andriessen, & Kanselaar, 2007).

Argumentation tools constitute a specific type of ERs. They are usually embedded in computer-supported collaborative learning (CSCL) environments and are used to stimulate students to explicate their claims and arguments. Students can often add arguments, counter-arguments, evidence, and so on to the tool. Furthermore, students can draw relationships between elements of the diagram such as causality and dependence relationships. There is evidence that representational guidance influences the effects that argumentation tools have on students' reasoning, use and quality of argumentation, and learning (e.g., Schwarz et al., 2003; Suthers & Hundhausen, 2003; Van Drie et al., 2005). It is important to note that the effects of representational guidance can differ depending on the specific circumstances (e.g., the kind of task to be performed). A given kind of guidance may for example be positive for one kind of task and negative for another. Compared to an Outline tool, the Diagram used by Munneke et al. (2007) for example, did not stimulate deeper argumentation during chat discussions, but helped students make more elaborate representations.

Some researchers maintain that more research is needed to determine the optimal representational guidance for students' co-construction of ERs (Erkens, Jaspers, Prangsma, & Kanselaar, 2005; Suthers, 2001; Toth et al., 2002). Furthermore, since most research has tended to focus on (a) the argumentative processes students engage in while co-constructing these ERs, and (b) the quality of the co-constructed ER, little is known about the effects that representational guidance may have on individual learning outcomes (Suthers, 2003). The aim of this research is therefore to investigate the role of representational guidance by comparing the effects of two representational tools that differ with respect to the guidance they offer on (a) the quality of the representations constructed, (b) the quality of the group products (an essay), (c) students' performance on a knowledge post-test, (d) students' online collaborative activities, (e) students' perceptions of the usefulness of using the representation, and (f) students' perceptions of the collaborative process.

Before describing the research questions and method of research, we describe the potentials and pitfalls of the co-construction of ERs in more detail. Furthermore, we describe the two different representational tools (offering different representational guidance) used in this study, and explain why we expect representational guidance to have a positive effect on performance, learning, collaboration, and perceptions.

CO-CONSTRUCTION OF ERS: POTENTIALS AND PITFALLS

When students use argumentation tools during collaboration, this combines the construction of ERs with CSCL. Although both approaches seem promising and may complement each other, each has its own pitfalls. In this section we describe the potentials and pitfalls of constructing ERs, how construction of ERs and CSCL can complement each other, and how co-construction of ERs in CSCL environments also has its own pitfalls.

Potentials and Pitfalls of Constructing ERs

The *benefits* of constructing ERs have been ascribed to several different causes. First, several authors have pointed to the fact that processing and constructing ERs stimulates students to engage in mental activities that facilitate learning and transfer. The construction of ERs helps students refine their understanding of complex phenomena, because it stimulates them to engage in the processes of self-explanation, externalization, elicitation, and elaboration of knowledge (e.g., Cox, 1999; Fischer, Bruhn, Gräsel, & Mandl, 2002). Furthermore, constructing ERs can help students reach higher levels of abstraction (Ainsworth, 2006; Schwartz, 1995) and thus deeper understanding of the subject matter. ERs can also help students engage in types of reasoning that are beneficial for their understanding of the subject matter (Cox). Van Drie et al. (2005), for example, found an effect of the type of representation constructed on students' use of historical reasoning: matrices elicited more talk about historical change than did lists or diagrams. Finally, ERs can help students think about concepts and theories on a deeper level by focusing them on the important concepts and relationships in the domain of study (Quintana et al., 2004).

Second, ERs can also reduce the cognitive effort needed to solve a complex problem. Argumentation tools for example, can help students determine whether their argumentation is balanced and whether it is supported by evidence (Munneke et al., 2007) since there is a pictorial trace of arguments, supports, and counterarguments. In this way, ERs can support computational offloading (Ainsworth, 2006): students need to invest less effort to determine whether their argumentation is balanced (i.e., they can see it in the diagram), and they can thus devote their cognitive resources to other issues. Similarly, ERs can direct attention to unsolved parts of the problem (e.g., students notice they have mostly addressed arguments in favor of a position and have neglected counter arguments). As such, an ER conveys students' progress through the problem (Cox, 1999; Suthers, 2001). Students can use this information to their benefit, for example by deciding whether they need to work more on a specific aspect of the representation. Finally, the information in an ER may serve as an extended working memory, helping students to recall their knowledge and ideas more easily (Löhner, Van Joolingen, & Savelsbergh, 2003; Van Bruggen et al., 2002).

The construction of ERs can have *drawbacks* as well. One drawback is that there is always the possibility that students construct incorrect representations which foster misconceptions (Ainsworth, 2006). Furthermore, problems may arise when students misinterpret the format of the ER or when there is a mismatch between the ER and the students' own conceptual representation. An example of this might be when students interpret the meaning of an arrow in a diagram incorrectly (e.g., as a parent-child relationship as opposed to a causal relationship) and this subsequently results in the construction of a less than optimal representation. This is not necessarily a failure of the students, but it could also stem from the tool (e.g., the notation used in the tool is unfamiliar and thus students interpret it differently than was intended). Moreover, when students interpret a representation differently than the developer, it may still lead to a productive discussion. Prior knowledge may also play a role here. When students lack prior knowledge, they may have more difficulties comprehending and successfully completing ERs (Cox, 1999).

A final drawback of construction ERs concerns the cognitive load it requires. The cognitive demand imposed by the construction of ERs might become too high, thus preventing learning (Van Bruggen et al., 2002). For example, when students are asked to construct a representation of a topic, they have to find and process relevant information, place this information correctly in the representation by relating it to other information units, and take into account the representational format. This task may be so complex for learners that it prevents them from learning about the subject or may create extraneous cognitive load that is deleterious to learning (Kirschner, Sweller, & Clark, 2006).

How CSCL and Construction of ERs Can Complement Each Other

CSCL offers students an environment in which their tasks and communication can be supported by specific tools. Several researchers suggest that embedding tools in CSCL environments that allow the co-construction of ERs can *enhance the effects of both constructing ERs and CSCL* (Munneke et al., 2007; Suthers & Hundhausen, 2003). First, co-construction of an ER in a CSCL environment stimulates further cognitive processing of the learning material. Compared to individually constructing an ER, students are more stimulated to externalize their thinking, to ask questions, give clarifications, and engage in a process of interactive argumentation during co-construction of ERs (Chinn & Anderson, 1998; Teasley

& Roschelle, 1993). These processes have been shown to be beneficial to students' learning processes. Second, although ERs themselves are able to trigger and support cognitive and social activities (Suthers, 2001), CSCL environments offer tools that can further support these activities. Some environments for example, offer tools that stimulate careful planning or reflective activities (e.g., Erkens et al., 2005). Other environments try to support the social process by scripting the collaborative process (e.g., Dillenbourg, 2002). These tools may facilitate the process of creating ERs. Furthermore, ERs can fulfill an important social and communicative function (Cox, 1999; Löhner et al., 2003) which can be capitalized on in a CSCL environment. Through co-construction of ERs, students represent their current ideas and conceptions about the problem. In their discussions students can refer to the ER, thereby helping them to create a common frame of reference (Suthers & Hundhausen, 2003). Through the co-construction of an ER and the discussion it creates, a joint problem space is created between learners and negotiation of meaning is facilitated (Erkens et al.; Teasley & Roschelle, 1993). Creating and sustaining such a joint problem space is a necessary prerequisite for successful collaboration.

Pitfalls of Co-constructing ERs in CSCL Environments

Unfortunately, research investigating the co-construction of ERs in CSCL environments shows that *this approach has its own disadvantages*. First, CSCL often requires students to communicate electronically. It is known that computer-mediated communication (CMC) can be difficult (Fjermestad, 2004). During CMC, group members often experience communication and collaboration problems. Students have difficulties interpreting group members' messages, for example due to the reduced number of cues (e.g., intonation of voice, gestures) available in such environments (Clark & Brennan, 1991; Daft & Lengel, 1986; Olson & Olson, 2000). Furthermore, research has often found that students have difficulties engaging in interactive argumentation (Kuhn & Udell, 2003; Munneke et al., 2007) leading to mixed outcomes of co-constructing ERs. Finally, the fact that students have to construct ERs together instead of individually may make an already challenging task even more difficult. Students have to coordinate their contributions to the representation. This means they have to make sure their partners understand their additions to the representation (Cox, 1999). Although explaining one's opinion or problem solution to group member can be beneficial to one's learning (cf., Webb & Palincsar, 1996), it also increases the cognitive demands placed on learners (Van Bruggen et al., 2002). If the cognitive demands become too high, chances are that students will not benefit from co-constructing ERs (Zhang, 1997). The question is whether the benefits of co-constructing ERs outweigh their demands. If this is not the case, students' learning processes may suffer.

Thus, co-constructing ERs is a promising educational approach, but has its own pitfalls. To help avoid these pitfalls, an important aspect to consider is the representational guidance an ER offers to students. Previous research has demonstrated that representations that differ with respect to the guidance they offer, stimulate different forms of reasoning and argumentation, and lead to different learning outcomes (Suthers & Hundhausen, 2003; Van Drie et al., 2005). In the next section we will therefore introduce two versions of a tool used for the construction of ERs in the domain of history. We will describe why these tools can have different effects on the process of co-constructing ERs.

CO-CONSTRUCTING A HISTORICAL DEBATE USING THE DEBATE-TOOL

This study involves students co-constructing an ER in a CSCL environment about a topic in the domain of history. Group members (group sizes range from two to four students) are given a number of historical sources and are asked to use the ER to collaboratively construct a representation of the topic. This activity precedes a writing task where students have to co-author an essay on a specific historical topic that includes their findings. To aid them in this process the Debate-tool was developed. In this section we describe two different versions of the Debate-tool, which differed in the representational guidance they offered to the students.

An argument or debate involves participants taking different positions on a topic or position. This also applies to historical debates. In this study, students studied the first centuries of Christianity, and more particularly the way Christians were treated or maltreated in the Roman Empire. There is some debate concerning the magnitude of these persecutions. On the one hand, some historians and the Catholic Church maintain that the persecutions were severe and a great number of Christians became martyrs because they were killed for religious reasons (*Martyrs position*). On the other hand, there is evidence this number was greatly and deliberately exaggerated by the church and Christians as propaganda meant to inspire other Christians and to increase the number of converts (*Propaganda position*).

In this study, students were given a number of historical and contemporary information sources and were asked to analyze and reconstruct this debate by co-constructing a representation of it with a Debate-tool and by collectively writing an argumentative essay based on their findings. In the Debate-tool and the essay, students were required to mention and discuss the arguments put forth by proponents of both positions. This required students to process and synthesize the information given in the sources since some arguments had to be constructed from information from multiple sources, because some sources contradicted each other, and because some sources contained archaic language and complex sentences. Thus, constructing an adequate representation required students to do more than simply cut-and-paste information from the sources to the tool.

Figures 1 and 2 show screenshots of both versions of the Debate-tool. We will refer to the version displayed in Figure 1 as the *Graphical Debate (GD)-tool* and to the version displayed in Figure 2 as the *Textual Debate (TD)-tool*. Both versions are used to reconstruct the historical debate. In Figure 1, the boxes labeled Martyrs and Propaganda on the edge of the inner circle represent both positions of the debate. While working with the GD-tool, students can add arguments (the boxes on the edge of the second circle) to either of the positions. These arguments can be found in, or inferred from the given sources. The sources also contain information that supports or refutes the arguments students add to the tool. This is represented by the boxes on the edge of the outer circle. Elements that represent supporting information have a white background, while elements that represent refuting information have a grey background. In the actual tool, these elements were color coded (i.e., supporting information using a green color and refuting information using a red color). The lines between the boxes indicate relationships between the elements of the representation.

In the TD-tool displayed in Figure 2, both positions are given as well. In this version of the tool, students also add arguments to the corresponding positions. No distinction is made however, between arguments, supports, and refutations. Instead, information is added to the TD-tool in a listwise manner (cf., Erkens et al., 2005; Van Drie et al., 2005).

--- INSERT FIGURES 1 AND 2 ABOUT HERE ---

In both versions of the tool, students can add elements by clicking on the Add-button. This opens a new window in which students can specify to which position the element will be added. In addition, they can give the new element a title (displayed in the tool), type in an explanation, and specify from which source the information came. In the GD-tool students can also specify whether the new element is an argument, a support, or a refutation. The process of co-constructing representations (i.e., reading and processing historical sources, extracting relevant information, placing this information in the appropriate place in the representation) is almost the same for both versions of the Debate-tool.

The biggest difference between the GD- and the TD-tool concerns the representational guidance they offer. Compared to the TD-tool, the GD-tool uses several visualization techniques to make information salient and to help students complete the representation more effectively and efficiently. These techniques include making the strength of support of a position visible through automatic repositioning of a position and accentuation of the complexity of the argumentation. These techniques will be discussed in the next paragraphs.

A first, rather obvious, difference between the two versions concerns their modality. While the TD-tool is mostly text based, the GD-tool uses both visual and textual elements to represent information.

Secondly, the GD-tool discerns between arguments, supports, and refutations. This feature may guide students to find supporting and refuting information, and to formulate arguments since it is immediately clear to them if this information is present or not. Furthermore, previous research has shown that students often have difficulties finding and formulating counterarguments (e.g., Kuhn & Udell, 2003). The option to distinguish between arguments, supports, and refutations may stimulate students to search for counterevidence. Furthermore, by guiding students in this direction, students may also be more likely to formulate arguments, supports, and refutations in their essays. On the other hand, it could be argued that, because of its increased complexity, the GD-tool is less suited to support the linearization process that occurs when writers have to transfer the content of the ER to an essay.

Third, the GD-tool visualizes how well positions are supported by arguments and supporting information. Each time an argument is added to a position, it moves two steps closer to the central flag. Moreover, whenever a supporting piece of evidence is added to an argument, the position moves one step closer to the flag. Finally, when a refutation is added to an argument, the position moves one step away from the flag. Consequently, when a position is located closer to the flag, it is better supported by arguments: the argumentation is more strongly in favor of the position. In Figure 1, the Propaganda position is located nearer to the flag than is the Martyrs position, thus in its current form the Propaganda position is better supported with arguments and supporting information. It is more difficult to infer this from the TD-tool because no distinction is made between arguments, supports, and refutations, although the lengths of the lists corresponding to the positions do give clues of this. The embedded representational guidance of the GD-tool may help students to draw a conclusion (Which position is supported best by arguments and information?), and thus may contribute to computational offloading (Ainsworth, 2006).

Fourth, the GD-tool visualizes students' progress through the problem (Cox, 1999) better than does the TD-tool. Although both versions give an idea about the balance between both positions (e.g., Has one position been given more attention than the other?) the GD-tool makes it clearer whether students tended to focus on certain elements (i.e., arguments, supports, or refutations). Students can infer this from the number and type of boxes that have been added to the representation. Furthermore, the boldness of the lines around the position and argument boxes also serves as an indication for their elaborateness and complexity. For example, in Figure 1, the lines around the Propaganda box are bolder compared to the lines around the Martyrs box, indicating the Propaganda position is more elaborate in terms of number of arguments, supports, and refutations. This may thus draw group members' attention to unaddressed parts of the problem.

Finally, in the GD-tool students have the option to rate the quality of arguments, supports, and refutations. Sometimes an argument may be more important or valuable than another argument. Students can express this by giving ratings to arguments, positions, and refutations. In Figure 1, the elements entitled ‘Voluntary martyrs’, ‘Emperor Trajan’, and ‘Exaggerated’ received a rating once (indicated by the star in the corresponding boxes), while the other elements did not receive a rating. This rating indicates the students thought these elements were more important than other elements. A rating influences the distance of the position from the flag. When a rating is given to an argument or support, its corresponding position moves one extra step closer to the flag. On the other hand, when a rating is given to a refutation, the corresponding position moves one extra step away from the flag. Students can rate an argument more than once, to further differentiate the importance of elements. The total of number of ratings that can be given by the group is limited (i.e., dependent on the size of the group). The rating functionality of the GD-tool serves two purposes. It stimulates students to think about and discuss the importance of arguments and may help them to see which arguments are more important than others.

RESEARCH QUESTIONS

The GD-tool offers different representational guidance to students compared to the TD-tool by using several visualization techniques. Aim of this study is to investigate whether this affects the quality of the representations students’ construct, the essays they write, their learning as measured by their performance on a knowledge test, their collaborative activities, and their perceptions of the tool as their online collaboration and communication. In sum, the following research questions are addressed: Do students who work with the GD-tool (GD condition), compared to students who work with the TD-tool (TD condition), …

1. … construct representations of higher quality?
2. … write essays of higher quality in terms of (a) how well all the topics in the given sources are covered, (b) quality of the grounds used in the essays, and (c) conceptual quality of the arguments used in the essays?
3. … perform better on a knowledge post-test?
4. … collaborate in a different way online?
5. … perceive a higher usefulness of the Debate-tool?
6. … perceive their online collaboration and communication more positively?

METHOD AND INSTRUMENTATION

Participants

The participants were students from five different history classes from two secondary schools. The total sample consisted of 124 eleventh-grade students (55 male, 69 female), with an average age of 16.24 years of age (*SD* = 0.57). Their teachers randomly assigned them to different groups. Due to uneven class sizes and student drop-out, this resulted in one 2-person group, 30 3-person groups, and eight 4-person groups. Because the 2- and 4-person groups collaborated in the same manner as the 3-person groups, they were not removed from the analyses.

Design

To investigate the effects of the Debate-tool, we used a single-factor, between subjects design with two different groups defined by the type of representation used: GD-tool or TD-tool. We randomly assigned three classes to the GD condition, and two classes to the TD condition. Classes rather than groups were chosen as the unit of assignment, because having two versions of the tool in the same class was considered undesirable (e.g., students would notice some groups had another version of the tool and this might influence their behavior). In total, 79 students in 24 groups (17 3-person and 7 4-person groups) worked in the GD condition, and 45 students in 15 groups (1 2-person, 13 3-person, and 1 4-person groups) formed the TD condition.

Task and Materials

CSCL Environment: VCRI

Students worked in a CSCL environment named *Virtual Collaborative Research Institute*. The VCRI is a groupware program developed to support collaborative learning on inquiry tasks and is composed of several separate, but interrelated tools. Students use the *Chat* tool to synchronously communicate with other group members. To read the description of their group task or to search and read relevant information, students can use the *Sources* tool. This tool lists a number of sources which can be opened and read from the screen. Group members use the *Cowriter* as a shared word processor. Using the *Cowriter*, group members can simultaneously work on different parts of their texts. To collaboratively construct (argumentative) diagrams, students can use the *Diagrammer*. VCRI contains several other tools designed to support for example the planning of the task (*Planner*), and the monitoring of group members' participation during the collaborative process (*Participation*-tool, see Janssen, Erkens, Kanselaar, & Jaspers, 2007). While working with

VCRI, each student works on a separate computer (i.e., students do not share their computer). This is done to ensure students do not divide their work (i.e., one student types on the keyboard, while the other uses the mouse) and that one student is less easily able to dictate was is added to the environment. Both issues have for example been documented when students share one computer to make a representation in Belvedere (cf., Suthers & Weiner, 1995).

--- INSERT FIGURE 3 ABOUT HERE ---

Task

Students collaborated on an inquiry group task in the domain of history. Topic of the task was “The first four centuries of Christianity” (Janssen, Erkens, & Kanselaar, 2007). During the first part of the task, students were given 14 historical and contemporary information sources and were asked to explore and discuss the different sources with respect to the Martyrs versus Propaganda debate. Students were required to co-construct a representation of this debate in either the GD- or TD-tool. After they had completed their representation, they had to co-author an argumentative essay based on their findings. Students had three lessons of 50 minutes each to complete this part of the task and submit their products to their teachers. The quality of the representations constructed and the essays written during this part of the task were analyzed to answer research questions 1 and 2 (see below). This part of the task can be characterized as an open and complex task. The 14 sources were quite difficult and students often had to synthesize information from multiple sources to construct arguments that pertained to one of the two positions. Furthermore, information in some sources contradicted information in other sources. This required students to discuss, reason, and argue about the information in the sources, the arguments they presented in the ER, and the information they included in their essays.

To complete the second part of the group task, the groups had to study 25 new sources about the topic. These sources needed to be categorized into up to five different categories by the group members. Students needed to decide on these categories themselves. Furthermore, group members were instructed to construct a diagram of their categorization using the Diagrammer. For the final part of the inquiry task, group members had to collaboratively write a second essay of at least 1200 words explaining why and how Christianity developed from a small ‘cult’ into the main religion of the Roman Empire.

It is important to note that parts 2 and 3 of the inquiry task were the same for both conditions. Furthermore, the instruction and task for part 1 was also the same for both conditions. The only difference concerned the explanation that students had to use either the GD- or TD-tool to construct a representation of the historical debate in part one of the inquiry task.

Research Question 1: Quality of Representations

To determine whether groups in the GD condition constructed representations of higher quality than the groups in the TD condition, we rated all of the items (arguments, supports, and refutations) placed in the tool on a 5-point scale (ranging from 0 - 4). To facilitate the rating of the items, we developed a flowchart based on the work of Clark and Sampson (2005) that contained a series of yes/no questions. These questions addressed quality aspects, such as (a) appropriateness of the item for the position that it was supposed to represent, (b) correctness and elaborateness of the explanation given to justify its inclusion, (c) use of additional information to mitigate or counter the main argument, and (d) use of references to historical persons, eras, or events to explain the historical context. Based on the answers given, we gave a score to each item, such that items that were clearly not appropriate for the position they were supposed to represent received a rating of 0. We gave a rating of 4 to items that contained an elaborate explanation, information to nuance or counter the main argument, and a historical context. Sometimes however, the items contained information that could be considered a repetition or paraphrase of information that had already been given in an earlier item. If this was the case, we marked the item to be a repetition.

Because some groups constructed representations with more items than others, we calculated a *mean quality score* for each group. If a group's representation contained repetitions of a topic, we only included the item that had received the highest rating in the mean quality score. Otherwise groups that had often repeated themselves several times would have an advantage over groups that used fewer repetitions. Furthermore, we also calculated the total number of items included in the representation, to check whether the GD-tool had an effect on the verbosity of the constructed representation.

Two raters (the first and second author) independently rated the quality of 85 Debate-items from four different groups. The agreement percentage between the two raters was 76%, while Cohen's *κ* was .69. The five categories of the rating scheme can be considered to represent an order (with 0 being the lowest, and 4 being the highest category), and thus a weighted *κ* was also calculated (Cohen, 1968; Kenny, Kashy, & Cook, 2006). The weighted *κ* takes the size of the discrepancies between raters into account, meaning that if rater A gives a rating of 0 to a certain item, while rater B gives a rating of 4, this discrepancy is given more weight than if rater A gives a rating of 2, while rating B gives a rating of 3. The weighted *κ* was found to be .73. When evaluating the found *κ* and weighted *κ* against Krippendorff's (1980) standard of .70, we concluded that the rating procedure was sufficiently reliable.

Research Question 2: Quality of Essays

After constructing a representation of the two positions in their Debate-tools, the students collaboratively wrote an essay detailing their findings. To determine whether groups in the GD condition wrote better essays than groups in the TD condition, we analyzed the quality of these essays with respect to (a) how well the essay covered all the topics that could be found in the historical sources, (b) quality of grounds used, and (c) conceptual quality of the argumentation.

Topics Covered in the Essays

To determine how well the essays covered all topics presented in the sources, we had to segment them into smaller units since each text contained several topics. Based on a detailed analysis of the historical sources available to the students, we distinguished 12 topics that related to the Propaganda position, and 6 topics that related to the Martyrs position. A graphical representation of the 18 different topics and the relationships between them is given in Figure 4. These topics were the starting point for the segmentation process. During the development of the segmentation process, we noted that students used two additional “topics” in their texts, which could not be found in the historical sources. The first topic (O1) concerned the assessment of arguments or positions and the drawing of conclusions (e.g., “The Propaganda position is supported better by arguments”). The second topic (O2) concerned text fragments, such as introductory sentences, which did not belong to any of the other topics (e.g., “This essay describes our findings on the Martyrs versus Propaganda debate”).

Segmentation involved reading each sentence and determining whether it signaled the beginning of a new topic. If this was the case, it also marked the end of the previous segment. After segmentation, we assigned a code to each segment corresponding to its topic. Examples of five segments and their topics can be found in Table 1. After the segmentation and coding of all the topics, we determined the total number of different topics covered for each group.

--- INSERT FIGURE 4 AND TABLE 1 ABOUT HERE ---

Grounds and Conceptual Quality of Segments

The distinction between *grounds quality* (how well students used evidence or examples to support their claims) and *conceptual quality* (how conceptually correct the arguments were) is made by several authors (e.g., Clark, Sampson, Weinberger, & Erkens, 2007; Munneke et al., 2007). These authors argue that examining grounds quality gives important information about how well students develop and support their arguments with evidence and explanations, but provides no details about the validity and conceptual quality of these arguments. It is possible that students receive a high score in terms of grounds quality when they write an argument that contains several explanations and includes multiple pieces of evidence. However, at the same time the conceptual quality of this argument may be lower, because the evidence students include or the explanations they give are (partially) flawed or incorrect (e.g., examples 1 and 4 in Table 1). Thus, we decided to code each segment in terms of grounds quality as well as conceptual quality.

Grounds quality. The evidence provided by students to back up the claims and opinions in their texts formed the starting point for the analyses of grounds quality. Each text segment was judged in terms of how well and how elaborately it was supported by evidence or explanations (Clark et al., 2007). Grounds quality of each segment was rated on a 4-point scale, ranging from 0 to 3. Segments that included no grounds in the form of explanations or evidence received a score of 0 (e.g., example 3 in Table 1). When only a short explanation was given to support the main claim of the segment, a score of 1 was given (e.g., example 5).

Segments that included an elaborate explanation or example to support the claim of the segment, received a score of 2 (e.g., example 1). Finally, segments that included an elaborate explanation or example with one or more references to historical sources were given a score of 3 (e.g., examples 2 and 4). This analysis was not applied to segments assigned to one of the two "Other" topics (O1 and O2).

Conceptual quality. The conceptual adequacy of the arguments given by the students, constituted the basis for the analyses of conceptual quality (Clark et al., 2007). Each segment was judged in terms of its conceptual correctness; thus segments that contained, for example, flawed conclusions, misinterpretations, or incorrect statements, received lower scores for conceptual quality than did segments that contained no errors. Conceptual quality was also rated on a 4-point scale (0 - 3). Segments that consisted of erroneous statements received a score of 0. Segments that contained some incorrect statements but also some correct ones received a score of 1 (e.g., examples 1, 3, and 5 in Table 1), whereas segments that contained statements that were mainly correct but that were sometimes a bit oversimplified received a score of 2 (e.g., example 4). Finally, segments that contained only conceptually correct and valid statements received a score of 3 (e.g., example 2).

As some groups covered more topics in their essays than others, we calculated the total grounds and conceptual quality as well as the mean grounds and mean conceptual quality for each group. These scores were used in the comparisons between the GD and TD conditions. Furthermore, we also calculated, for each group, the mean length in number of words of the topics. This was done to check whether groups that wrote larger, more complete topics received higher scores for essay quality, and whether this possibly interacted with an effect of our two conditions.

Reliability Analyses

The first author and a research assistant were involved the reliability analyses of the segmentation procedure, topic analysis, and analyses of grounds and conceptual quality (see Table 2). The reliability analysis of the *segmentation procedure* involved a comparison of the boundaries of the segments as determined by the two researchers. Whenever these boundaries differed between the two researchers, this was counted as a disagreement. Because the researchers could differ in the number of segments they discerned, the percentage of agreement was determined from the perspective of both researchers. This gave an upper and a lower bound agreement percentage (see Strijbos, Martens, Prins, & Jochems, 2006). Agreement percentage, Cohen's *κ*, and category kappa's (Cicchetti, Lee, Fontana, & Dowds, 1978) were calculated to judge the objectivity of the topic analysis. Besides agreement percentage, Cohen's *κ*, and category kappa's, a weighted Cohen's *κ* was also calculated for grounds quality and conceptual quality. The moderate to high reliability coefficients shown in Table 2 indicate the analysis of the quality of the written essays was sufficiently reliable.

--- INSERT TABLE 2 ABOUT HERE ---

Research Question 3: Pre- and Post-test Knowledge Measures

To investigate whether students in the GD condition learned more compared to those who worked in the TD condition, knowledge pre- and post-tests were developed. Both tests consisted of the same 15 multiple-choice items addressing topics covered in the inquiry group task. The test contained factual (Why did Christians refuse to worship the Roman Emperor?) as well as comprehension or insight questions (Explaining why a historical source was written by a Roman or a Christian). We used Cronbach's alpha to determine test reliability. Pre-test reliability was .51. This indicates low item homogeneity, probably due to the fact that students had low to average prior knowledge, causing them to guess the correct answer. Post-test reliability was .67, which can be considered acceptable.

Research Question 4: Online Collaboration

There is evidence that representational guidance can have an impact on the discussions that unfold between group members as they construct a representation. Van Drie et al. (2005) for example, found that students used more historical reasoning in their discussions when they used a Matrix representation compared to students who constructed a Diagram or List. Furthermore, the representation also had an effect on regulative and coordinative processes: matrix users discussed more about how they were going to approach the task. There is therefore reason to believe students in the GD condition will collaborate differently than students in the TD condition. For example, the GD-tool offers more information about students' progress through the task (i.e., Have both positions been given equal attention?) and thus might decrease the need for active monitoring of task-progress. To examine the effect of representational guidance on students' collaborative activities, we analyzed their online collaboration with a coding scheme (see Janssen, Erkens, & Kanselaar, 2007; Janssen, Erkens, Kanselaar et al., 2007).

When students collaborate on an inquiry task, they need to exchange their ideas and opinions, ask questions, and work towards producing a group product (McGrath, 1991). On the other hand, collaboration also involves a social-relational aspect. Students have to perform social and communicative activities that establish group well-being (Meier, Spada, & Rummel, 2007). Therefore, the coding scheme not only contains codes relating to content, but also several codes that refer to the social and communicative aspects of collaboration, such as greeting each other or engaging in activities that contribute a positive group climate (e.g., joking, social talk).

On the other hand, collaboration requires considerable coordination and regulation of these activities (Erkens et al., 2005; Meier et al., 2007). Metacognitive activities that regulate task performance (e.g., making plans, monitoring task progress, and evaluating plans or ideas) are considered important for online collaboration (De Jong, Kollöffel, Van der Meijden, Kleine Staarman, & Janssen, 2005). Moreover, social activities have to be coordinated and regulated as well (Manlove, Lazonder, & De Jong, 2006). For instance, students have to discuss and plan their collaboration, monitor their collaboration, and evaluate their collaborative process. Thus, the coding scheme also contained codes that referred to the regulation and coordination of task-related and social activities.

In total, the scheme contains four dimensions: *task-related activities, regulation of task-related activities*, *social activities*, and *regulation of social activities*. Each dimension contains two or more codes or so-called collaborative activities. Furthermore, the scheme included several additional categories (e.g., technical aspects) that did not belong to any of the four dimensions. In total, the scheme consisted of 19 categories (see Table 3).

Segmentation and coding procedure. Before the start of the coding process, it was necessary to segment students' messages into smaller parts that were meaningful themselves. This was done because multiple concepts or ideas were often conveyed within one message. Therefore, the chat messages were segmented into smaller units.

Segmentation and coding were done using the *Multiple Episode Protocol Analysis* (MEPA) computer program (Erkens, 2005). Messages were segmented using a segmentation filter. A filter is a program, which can be used in MEPA for automatic data manipulation. Punctuation marks (e.g., full stop, exclamation mark, question mark, comma) and phrases connected through a conjunction (e.g., "and", "but", "or") are used to segment messages into dialogue acts (Erkens & Janssen, in press). Using filters speeds up segmentation, and ensures segmentation rules are applied consistently. After the segmentation process, the segments were coded using the coding scheme.

Interrater reliability. Two researchers were involved in the interrater reliability analysis. Ten protocols were chosen at random, and from each of these protocols 100 segments were randomly selected. Thus, in total 1,000 segments were independently coded by both researchers. The overall Cohen's κ for coding of collaborative activities was .90 (see Table 3).

--- INSERT TABLE 3 ABOUT HERE ---

Research Question 5: Perceived Usefulness of Debate-tool

To check whether students in the GD condition perceived higher usefulness of the tool compared to students in the TD condition, the post-test questionnaire contained a scale that tapped into their perceptions of its usefulness. The scale consisted of eight items rated on a five-point scale. Sample items of this scale are "The Debate-tool made it easy for us to write the essay" and "The Debate-tool gave us a good overview of the two positions." The eight items formed a homogeneous scale, as indicated by a Cronbach's alpha of .80.

Research Question 6: Perception of Online Communication and Collaboration

Students' perceptions of their online communication and collaboration were measured with three different scales in the post-test questionnaire. The first scale addressed *positive group behavior* and consisted of seven items rated on a 5-point scale. Behaviors such as helping each other and equal participation among group members are indications of positive group behavior (Webb & Palincsar, 1996). A sample item from this scale is: "We helped each other during collaboration." The second scale tapped into *negative group behavior* and consisted of five items rated on a 5-point scale. Conflicts and free riding behavior (O'Donnell & O'Kelly, 1994) are indications of negative group behavior. “There were conflicts in our group” is an example from this scale. The final scale addressed students’ *perceived effectiveness of their group’s task strategies*. This scale was based on the work of Saavedra, Early, and Van Dyne (1993) and consisted of eight items that assessed the choices made and the strategies chosen by the group members. Again, these items were rated on a 5-point scale. An example from this scale is: “We planned our group work effectively.” All three scales had adequate reliability coefficients (.83, .70, and .74 respectively).

Procedure

Before the start of the inquiry task, students completed the knowledge pre-test and received information about the task and the group compositions from their teachers. During the lessons, each student worked on a separate computer in a computer lab. Students were therefore often located in the same room during the lessons and could thus also talk face-to-face to each other. It should be noted however, that students mostly communicated electronically. The role of the teacher was similar in both conditions: they were online to answer task-related questions during all lessons. Students were also allowed to work on the inquiry group task during free periods. Thus, they could work on the task in the school’s media center when they had spare time in their timetable. After the third lesson, students submitted the final version of their representations in the Debate-tool and their essays for the first part of the task. These were subsequently used in the analyses conducted for research questions 1 and 2. After the eighth and final lesson the knowledge post-test and questionnaires on perceived usefulness of the Debate-tool and perception of online communication and collaboration were administered to the students.

Data Analyses

When analyzing the effects of representational guidance, the data-analytical problem of *nonindependence* had to be taken into account (Kenny et al., 2006). Because students worked in groups in this study, they influenced each other. This violates the assumption of nonindependence of observations of individuals, making the results of traditional analytical techniques unreliable. Multilevel analysis can cope with nonindependence and is therefore a more appropriate technique. Thus, in the case where individual level variables were used (research questions 3, 4, 5, and 6), multilevel analyses were conducted. However, when the group was the unit of analysis, nonindependence is not a problem and thus in these cases (research questions 1 and 2) analyses of variance were conducted.

RESULTS

Time Online and Tool Use

We analyzed students' online activities to rule out the possibility that differences between conditions concerning quality of representations, quality of the essays, and so on could be attributed to students being more active in one version of the Debate-tool than in the other. First, as noted earlier, students worked with the GD-tool for three 50-minute lessons. Because students could also work during their free periods, it was necessary to check whether the total time students were online was equal in both conditions. Furthermore, we calculated how often the following three types of activities were performed by the students: adding, editing, and deleting elements in the Debate-tool. Table 4 shows the descriptive statistics for these activities and total time online, distinguishing between the two conditions. No significant effects of condition on time online (β = -1.46, *p* = .16), adding (β = 0.00, *p* = .99), editing (β = 0.17, *p* = .80), and deleting elements (β = -0.10, *p* = .72) in the Debate-tool were found using multilevel analyses. Thus, GD students were not more active in their version of the tool than TD students.

--- INSERT TABLE 4 ABOUT HERE ---

Quality of Representations

Table 5 shows the average number of items included in the tools and the average quality of the representations for groups in the GD and TD condition. No significant difference was found between both groups with respect to the number of items they included in their representations. Groups in the GD condition however, made representations of significantly higher quality than groups in the TD condition. This applied to the Martyrs as well as the Propaganda position. The corresponding effect sizes (Cohen's *d*) were large. Additionally, we correlated the number of items produced with the average quality of these items and found a significantly negative correlation, *r* = -.66, *p* = .00, meaning that when groups attempted to included a large number of items in their representations, this had a negative effect on the quality of their representations.

--- INSERT TABLE 5 ABOUT HERE ---

Quality of Essays

Table 6 shows the means and standard deviations for GD and TD groups with respect to the number of topics covered, total grounds and conceptual quality, and average grounds and conceptual quality. Multivariate analyses of covariance were conducted to test the effects of condition (C), length of the topics (L), and the interaction between condition and length of the topics (C * L). As can be seen in Table 7, the multivariate tests of condition, length, and condition by length were all significant. The univariate analyses of covariance shown in Table 7 indicate that GD groups did not cover more topics in their essays than did TD groups, nor were their essays better in terms of average grounds quality. However, GD groups produced significantly better essays in terms of total grounds quality, total conceptual quality, and average conceptual quality.

Length of the topics was found to have an effect on average grounds quality and average conceptual quality. Inspection of the regression coefficients revealed that longer topics were associated with higher average grounds and conceptual quality.

Finally, statistically significant interactions between condition and length were found for total grounds quality and total conceptual quality. The regression coefficients indicate that when GD groups wrote short to average topics their total grounds and conceptual quality was significantly higher compared to TD groups. When GD groups wrote longer topics these differences disappeared, however.

Additional analyses revealed significantly positive correlations between total grounds quality and total conceptual quality and between average grounds quality and average conceptual quality (r = .89, *p* = .00, and *r* = .47, *p* = .00 respectively), meaning that groups that received high scores for grounds quality also received high scores for conceptual quality. Interestingly, a significantly negative correlation was found between the number of topics covered in the essays and average conceptual quality (*r* = -.34, *p* = .04). Thus, when groups attempted to cover a large number of topics in their essays, this had a negative effect on the conceptual quality of their essays.

--- INSERT TABLES 6 AND 7 ABOUT HERE ---

Post-test Performance

Unfortunately, on both measurement occasions several students were unable to complete the test due to various reasons. Most of these students (*N* = 29) were unable to complete the questionnaire because they were required to attend an information meeting for further education during one of the scheduled lessons for the pre- or post-test. A smaller group of students was ill (*N* = 11) or truant (*N* = 5) on one of these occasions. For the knowledge pre-test, test scores from 59 (75%) of the GD and 43 (96%) of the TD students were available. For the post-test, these numbers were 59 (75%) and 42 (93%) respectively. An analysis of the students with missing values shows that students that were unable to complete the post-test performed nearly as well on the pre-test (*M* = 10.88, *SD* = 1.87) as students that were able to complete the post-test (*M* = 11.31, *SD* = 2.23). Furthermore, students that were unable to complete the pre-test performed equally well on the post-test (*M* = 12.47, *SD* = 2.01) as students that were able to complete the pre-test (*M* = 12.76, *SD* = 2.05).

Descriptive statistics for GD and TD students' pre- and post-test performances are given in Table 8. As can be seen, student performance increased from pre- to post-test in both conditions. Multilevel analysis was used to answer the question whether students in the GD condition learned more compare to students in the TD condition. The first step in this analysis was to examine the results of a model without any independent variables, the so-called null model (Model 0, cf. Snijders & Bosker, 1999). As can be seen in Table 9, a considerable part (1.33 / (1.33 + 2.83) = .32) of the total variance in post-test scores can be attributed to group-level factors. This indicates that not only individual student characteristics such as prior knowledge and motivation, but also group characteristics such as nature and quality of the online collaboration and group composition have an effect on post-test performance.

To determine the effect of condition, while controlling for prior knowledge, condition and pre-test score were added to the model (Model 1). Both pre-test performance and condition had a significant effect on students' post-test performance. As expected, a higher pre-test score contributed to a better post-test performance. Furthermore, condition contributed significantly to post-test performance, indicating a positive effect of working with the GD-tool. As can be seen in Table 9, adding these two predictors resulted in a significant decrease in the deviance of the model, indicating a better model fit. To check whether the GD-tool had a differential effect for learners with low or high prior knowledge, we added a condition by pretest interaction term to the multilevel model. No indication for such a differential effect of the GD-tool was found, $\beta = 0.03$, $p = .39$; $\chi^2 = 0.07$, $df = 1$, $p = .40$.

--- INSERT TABLES 8 AND 9 ABOUT HERE ---

Online Collaboration

Table 10 shows the effect of condition on students' online collaborative activities. Out of the 19 activities, the experimental condition was found to have an effect on four collaborative activities. Condition had positive effects on *shared understanding* (SociUnd+), *negative social evaluations* (MSociEvl-), and *neutral technical remarks* (TechNeut), meaning that students in the GD condition engaged in these activities more than did students in the TDcondition. A negative effect of condition was found on *loss of shared understanding* (SociUnd-), meaning that GD students expressed less misunderstandings and disagreements than TD students.

--- INSERT TABLE 10 ABOUT HERE ---

Perceived Usefulness of Debate-tool

Overall, GD ($M = 3.32$, $SD = 0.62$) and TD students ($M = 3.20$, $SD = 0.67$) perceived their respective versions of the Debate-tool as useful. No effect of condition was found, $\beta = 0.06$, $p = .17$; $\chi^2 = 0.89$, $df = 1$, $p = .17$.

Perception of Online communication and Collaboration

On average, students in the GD condition ($M = 3.69$, $SD = 0.57$) reported lower levels of *positive group behavior* compared to students in the TD condition ($M = 3.91$, $SD = 0.51$). Indeed, multilevel analyses showed an effect of condition on perceived positive group behavior, $\beta = -0.11$, $p = .04$; $\chi^2 = 2.90$, $df = 1$, $p = .04$. This means working with the GD-tool had a negative effect on perceived positive group behavior. Concerning *negative group behavior*, students in the GD condition ($M = 2.39$, $SD = 0.69$) reported slightly higher levels than did students in the TD condition ($M = 2.26$, $SD = 0.58$). No significant effect of condition on perceived negative group behavior was found, $\beta = 0.07$, $p = .18$; $\chi^2 = 0.84$, $df = 1$, $p = .18$. Finally, students in the GD condition ($M = 3.52$, $SD = 0.47$) reported somewhat lower *effectiveness of their group's task strategies* compared to students in the TD condition ($M = 3.68$, $SD = 0.54$). However, the effect of condition on perceived effectiveness of group task strategies was not significant, $\beta = -0.09$, $p = .07$; $\chi^2 = 2.17$, $df = 1$, $p = .07$.

CONCLUSION AND DISCUSSION

Aim of this study was to examine the effects of representational guidance on students' performance of a group task, learning of historical concepts and facts, collaborative activities, and perceptions of each other, their collaboration, and of the tools used.

The first research question addressed the *quality of the representations* made by the participating groups. Our analyses show that groups working with the Graphical Debate (GD)-tool constructed external representations (ERs) of higher quality than groups working with the Textual Debate (TD)-tool. An explanation for this finding may lie in the cognitive effort students needed to devote to the construction of the ER. Because the GD-tool offered more information about the balance of the argumentation and unsolved parts of the problem than the TD-tool, this may have made the students' task relatively easier (Cox, 1999). The GD-tool offered students representational guidance, thus supporting computational offloading (Ainsworth, 2006). As a result, students may have been able to devote more effort to the conceptual content of their ERs. This finding is in line with previous research. Suthers and Hundhausen (2003) for example, found differences between Text, Graph, and Matrix representations with respect to number of hypotheses and evidential relations students expressed. Also, Van Drie et al. (2005) found effects of the type of representation on the number of arguments used, the balance of arguments for and against a position, and the number of historical sources used in the representation. Finally, Fischer et al. (2002) examined the differences between a content-specific and a content-unspecific representation. They found the solutions in the content-specific condition to be of higher quality than solutions in the content-unspecific condition. Our results therefore add to a growing body of research indicating the effects of representational guidance on the quality of the representations constructed

Our second research question addressed the *quality of the essays* written by the groups. We expected GD groups to outperform TD groups. Our analyses indeed showed an effect of the GD-tool on the average conceptual quality of students' essays, as well as on the total grounds and conceptual quality of these essays. The GD-tool was not found to have an effect on the number of topics covered and average grounds quality.

It is important to note that an interaction effect was found between condition and average length of the written essay topics. When groups wrote short to average sized topics, GD groups outperformed TD groups regarding total grounds and conceptual quality of their essays. When groups wrote relatively long topics, this effect disappeared. Apparently, the GD tool is more suited to improve the essay quality of groups that tend to write relatively verbose paragraphs or topics. Groups that write longer topics, may be groups that adopt a "more is better" strategy, wich may be an indication that they devote too much time trying to put as much information as possible into their essays without considering the quality of this information and the quality of their writing. This explanation is in line with the negative correlation we found between number of topics and essay quality. This finding will be discussed in more detail below.

Our finding of an effect of representational guidance on essay quality is somewhat different from other research findings, where limited effects of representational guidance on collaborative writing were found (e.g., Suthers & Hundhausen, 2003; Toth et al., 2002; Van Drie et al., 2005). An explanation may lie in the *representational guidance* offered by the GD-tool compared to the guidance offered by the tools in the work of other researchers. Our tool directs students' attention to the distinction between arguments, supports, and refutations, and this may stimulate students to incorporate these elements in their ERs and essays. It has been argued that tools that support *linearization*, that is the ordering of content and arguments into an essay, may be better supported by ERs specifically designed to support the planning of the linear structure of essays (e.g., an Outline tool as described in Erkens et al., 2005). Although the GD was not specifically designed to support the process of linearization, it may be the case that stimulating students to systematically address all arguments, supports, and refutations of a position also facilitates the process of converting a representation into an essay.

Interestingly, we found a negative correlation between the number of topics the groups covered in their essays and the conceptual quality of their essays. When groups tried to cover a large number of topics in their essays, this negatively affected the conceptual quality of their essays. An explanation might be that when groups try to cover a lot of topics in their essays, this limits the time and effort they can devote to ensuring their claims are conceptually correct. Furthermore, this also makes writing an essay more complex, further decreasing the cognitive resources students can devote to writing conceptually correct claims. The more complex and elaborate groups' essays become, the harder it will be for them to clearly reason and structure the essay. There might thus be a trade-off between completeness and the quality of the essay. Also, when groups try to address many topics (i.e., they adopt a more is better strategy), this constrains the time available to check the correctness of their claims and to create a clearly structured essay. Indeed, groups needed to work very hard to complete the task in the allotted time (3 lessons of 50 minutes each for this subtask) and thus there was quite some time pressure (although students could also work during their free periods). Considering the complexity of the task and the existing time pressure, trying to include too many topics in the essay may not have been an adequate strategy. A similar negative correlation was also found between number items and quality of the representation. This further points to the possibly negative effects of a more is better strategy.

We expected students working with the GD-tool to outperform students working with the TD-tool on a *knowledge post-test*. This expectation was confirmed. Additionally, it is important to note that students in both conditions improved their performance from pre- to post-test. Unfortunately, the number of students that were unable to complete either the pre-test, the post-test, or both was higher in the GD condition (25%) than in the TD condition (6%). This may have negatively affected the comparability of both conditions, although missing value analyses indicated no systematic effects of missing data on pre- or post-test performance.

Previous studies have reported mixed effects of representational guidance on students' post-test performance. Van Drie et al. (2005) found differences on post-test performance between different representations, while Suthers and Hundhausen (2003) did not. An explanation for this finding may lie in the time students' spent working on the representations. In this study, students worked for three 50-minute lessons with the GD- or TD-tool. Moreover, they worked another five lessons on the other two parts of the inquiry task. Similarly, in the Van Drie et al. study, students worked in the collaborative environment during six 50-minute lessons. In contrast, in the Suthers and Hundhausen study for example, students worked on their representations for only one session of approximately 45 minutes. Suthers and Hundhausen hypothesized that this might not have been sufficient for learning outcomes to fully develop. This may explain why in our and the Van Drie et al. study, differences between representations were found, while these were not found in the studies by Suthers and Hundhausen. In our case, students possibly worked long enough in the CSCL environments for learning outcomes to develop. Moreover, differences in learning outcomes between the GD- and TD-tool also had sufficient time to develop. This difference in time on task may also explain why in our study effects of representational guidance on the quality of essays were found and not in some other studies. When students are given more time to work on their representations, the effects of representational guidance on learning outcomes may be more profound.

Another explanation may lie in the type of problem used. In this study, the goal was to provide a complete as possible representation of the historical debate. This may be different from what is normally understood under argumentation, namely to persuade someone of one's own point of view. These different goals may lead to a different approach of the task by students and subsequently to different learning outcomes. For example, when Nussbaum (2005) instructed students to generate as many reasons as possible, they engaged more, and more deeply in argumentation than students who were instructed to persuade someone. Thus goal instruction may be another important factor to consider when studying the effects of representational guidance.

In previous research, representational guidance has been found to affect students' *collaborative activities*. However, guidance was only found to have an effect on 4 of the 19 collaborative activities investigated. The GD-tool was found to have an impact on *shared understanding* during collaboration, as students' discussions showed more indications of shared understanding, and less indications of loss of shared understanding. Furthermore, these groups also typed more *negative evaluations of the social process* and made more *neutral technical remarks*. This may be an indication that collaborating while using the GD-tool is more complex. Furthermore, we found no effect of condition on metacognitive activities (i.e., planning, monitoring, and evaluating the task process). Therefore our study offers no support for the expectation that representational guidance decreases group members' need to coordinate and regulate their task performance in the online discussions. This finding may be explained by the fact that group members had access to other tools that support metacognitive activities. In other studies (e.g., Suthers & Hundhausen, 2003), participants only worked with a representational tool and not with other tools and therefore the effects on metacognitive activities more profound. Moreover, the GD-tool did not trigger students to question each other more or to generate more ideas about the topic. This last finding may be explained by the fact that students could use the representations in both the GD-tool and the TD-tool to exchange information (Van Amelsvoort, Andriessen, & Kanselaar, 2008; Van Drie et al., 2005). Because both tools were shared, adding an element to the representation equates to exchanging information with group members. Thus, there might be less need to engage in extensive information exchange in the chat discussions.

Our last two research questions addressed the impact of representational guidance on *students' perceptions*. Contrary to our expectations, guidance did not have an impact on students' perceived usefulness of the Debate-tool. Students in both conditions perceived their tool as useful. When looking at group and student performance, these results are not surprising. In both conditions, groups constructed representations and wrote essays of at least average quality and student performance increased from pre- to post-test. In this sense, both representations were useful to the students. Furthermore, neither tool was designed to be aversive, thus equal levels of perceived usefulness in both conditions were perhaps to be expected. Also, students were not able to compare the two versions of the Debate-tool directly, because they only worked with one version of the tool. Therefore, we were not able to ask which tool they thought was more useful, but were forced to ask them to rate the usefulness of the version of the tool they used. Finally, one can question the validity of measures that ask students to rate the usefulness of something. Often students' perceptions of the usefulness of something do not correspond to its effectivity or efficiency. Students may be more concerned with 'getting the job done', than with constructing the best possible representation and engaging in argumentative discussion. Thus, what is useful for students may be very different from what we as educational designers consider useful (e.g., a tool that enables them to complete the task as fast as possible versus a tool that triggers deep and mindful study of the topic at hand).

We also expected the GD-tool to have positive effects on students' perceptions of their online communication and collaboration. This was not confirmed. On the contrary, students in the TD condition reported significantly higher levels of positive behavior. However, the negative effect of condition on perceived positive group behavior should not be interpreted to mean that GD students did not enjoy their collaborative experience. On a scale from 1 to 5, students in the GD condition rated positive group behavior with an average of 3.69. This indicates they too experienced the collaborative experience positively. On the other hand, when we examined students' online collaborative activities, we found they more often voiced negative evaluations of the collaborative process and that they made more remarks about technical aspects of the environment. Possibly, students experienced the GD-tool as a more difficult and complex tool for collaboration. Furthermore, the TD-tool may be more familiar to students because it resembles a threaded discussion, making it easier and more comfortable to work and collaborate with. This may explain why GD-students perceived their collaboration less positively. Furthermore, as stated above, students often approach argumentation in a quick and easy manner and are often aimed at maintaining consensus (Kuhn & Udell, 2003). Therefore, a tool which forces them to argue more elaborately and critically may lead to more negative feelings toward the collaborative process. This is sometimes called the *paradox of collaborative learning*: we assume students learn from elaborating, arguing, and criticizing, while students themselves are often aimed at reaching consensus to complete their task as fast and efficient as possible (Erkens, Prangsma, & Jaspers, 2006).

Several methodological issues should be kept in mind when interpreting the results of this study. In this study, we chose to assign whole classes instead of groups to our conditions (GD- or TD-tool). This was done because it was undesirable to have two versions of the tool in the same classroom (i.e., students would notice the differences and this might influence the results). On the other hand, this could have evoked a teacher effect: the teachers could have had an impact on the results due to differences in the way they guided their groups. This teacher effect could be analyzed using multilevel analysis. Our sample on the highest level (the class) is too small however ($N = 5$), to analyze class or teacher effects meaningfully.

When we did however analyze the data listed in Table 9 using three levels (class, group, individual) instead of two levels (group, individual) we found the same significant effects of pre-test score and condition. We interpret this as evidence that it is the condition that explains the significant effect on the post-test score and not the teacher. However, to be able to completely rule out any teacher effects, a larger sample employing more classes and teachers is needed.

This study was conducted to investigate how representational guidance affects co-construction of ERs and employed two conditions. In general, the results are in favor of the GD condition. However, it might be argued that our design does not permit us to conclude which elements of the GD-tool contributed to these positive effects. Was it the graphical nature of the tool? Was it the feedback about the strength of support for a position and the complexity of argumentation? Or was it a combination of both? To be able to answer these questions a more elaborate research design is needed. Future research might for example use a design with four conditions: (1) a graphical condition with feedback (i.e., the GD-tool), (2) a graphical condition without feedback, (3) a textual condition with feedback (i.e., accentuation of complexity of argumentation by increasing the font size), and (4) a textual condition without feedback (i.e., the TD-tool). Using such a design might provide evidence for which specific elements of representational guidance contribute to the co-construction of ERs.

Although this more elaborate experiment makes sense from a developer's or researcher's viewpoint, from a pedagogical viewpoint it makes less sense. Based on the results of this study, it could be argued that the *representational guidance as a whole* offered by the GD-tool was more beneficial than the guidance offered by the TD-tool. Why then offer learners versions of the Debate-tool that visualize less aspects of the problem solving process and run the risk their learning process becomes less optimal?

A final issue concerns the fact that students mostly worked in the same computer lab and could therefore talk to each other during the sessions and between sessions. Although most groups communicated quite intensively through the chat tool, we cannot entirely rule this possibility out. Of course, this may have influenced our results and may for example explain why we did not find an effect of the GD-tool on students' online collaboration. Furthermore, one may also question the educational value of 'forcing' students to argue online, when they could also easily argue face-to-face. In our opinion, our approach has its own educational value. For example, the simultaneous sharing of a text document in the Cowriter enables students to really co-author their essay (i.e., they do not have to wait until the other group members have finished writing). We could have chosen to let students work together on the same computer, but this could also have its disadvantages. In this case for example, one student could claim the keyboard and mouse, thereby dominating the collaborative situation. When each student works with his or her own computer, the possibility that one student dominates the collaborative processes becomes smaller. This view is supported by research. Fjermestad (2004) for example, compared face-to-face versus computer-mediated communication in a meta-analysis and found more equal participation when students communicated through the computer.

In spite of these limitations, this study shows the advantage of the GD-tool over the TD-tool. We were able to demonstrate the effectiveness of the GD-tool in terms of helping groups of students to construct meaningful and correct representations and clearly argued and conceptually correct essays. Furthermore, while working with the GD-tool, students acquired more knowledge about the topic of study. These are promising and important results, since increasing group and individual achievement is an important hallmark for establishing the effectiveness of an educational design (Kirschner, Strijbos, Kreijns, & Beers, 2004). One direction may be to combine the use of ERs with instruction in proper forms of argumentation as secondary education students often lack the necessary skills to engage in argumentation (Kuhn & Udell, 2003; Nussbaum, 2005). Finally, the possible interaction between the goal of the task (e.g., to generate as much arguments as possible or to persuade) and representational guidance deserves further study. Whichever direction future research takes, the results of this study firmly point to the need for teachers and researchers to carefully consider the representational guidance they offer to students when using representation tasks in their teaching or research.

ACKNOWLEDGEMENTS

This study is part of the Computerized Representation of Coordination in Collaborative Learning (CRoCiCL) project. This project is funded by NWO, the Netherlands Organisation for Scientific Research under project number 411-02-121. Furthermore, the authors would like to thank Jos Jaspers and Marcel Broeken for their technical assistance, Ilse Mink for her help during the data collection, and Renske de Kleijn for her assistance during the data analyses.

REFERENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*, 183-198.

Chinn, C. A., & Anderson, R. C. (1998). The structure of discussions that promote reasoning. *Teachers College Record, 100*, 315-368.

Cicchetti, D. V., Lee, C., Fontana, A. F., & Dowds, B. N. (1978). A computer program for assessing specific category rater agreement for qualitative data. *Educational and Psychological Measurement, 38*, 805-813.

Clark, D. B., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review, 19*, 343-374.

Clark, D. B., & Sampson, V. D. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions. In T. Chan & D. D. Suthers (Eds.), *Computer-supported collaborative Learning 2005: The next 10 years!* (pp. 76-85). Taipei, Taiwan: Lawrence Erlbaum Associates.

Clark, H. H., & Brennan, S. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: American Psychological Association.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction, 9*, 343-363.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*, 554-571.

De Jong, F., Kollöffel, B., Van der Meijden, H., Kleine Staarman, J., & Janssen, J. (2005). Regulative processes in individual, 3D and computer supported cooperative learning contexts. *Computers in Human Behavior, 21*, 645-670.

Dillenbourg, P. (2002). Over-scripting CSCL: The risk of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL: Can we support CSCL?* (pp. 61-91). Heerlen, The Netherlands: Open Universiteit.

Erkens, G. (2005). Multiple Episode Protocol Analysis (MEPA). Version 4.10. Retrieved October 24, 2005 from http://edugate.fss.uu.nl/mepa/

Erkens, G., & Janssen, J. (in press). Automatic coding of online collaboration protocols. *International Journal of Computer Supported Collaborative Learning*.

Erkens, G., Jaspers, J., Prangsma, M., & Kanselaar, G. (2005). Coordination processes in computer supported collaborative writing. *Computers in Human Behavior, 21*, 463-486.

Erkens, G., Prangsma, M., & Jaspers, J. (2006). Planning and coordinating activities in collaborative learning. In A. M. O'Donnell, C. E. Hmelo-Silver, & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 233-263). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction, 12*, 213-232.

Fjermestad, J. (2004). An analysis of communication mode in group support systems research. *Decision Support Systems, 37*, 239-263.

Janssen, J., Erkens, G., & Kanselaar, G. (2007). Visualization of agreement and discussion processes during computer-supported collaborative learning. *Computers in Human Behavior, 23*, 1105-1125.

Janssen, J., Erkens, G., Kanselaar, G., & Jaspers, J. (2007). Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Computers & Education, 49*, 1037-1065.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York/London: The Guilford Press.

Kirschner, P. A., Strijbos, J. W., Kreijns, K., & Beers, P. J. (2004). Designing electronic collaborative learning environments. *Educational Technology Research and Development, 52*(3), 47-66.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75-86.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*, 1245-1260.

Löhner, S., Van Joolingen, W. R., & Savelsbergh, E. R. (2003). The effect of external representation on constructing computer models of complex phenomena. *Instructional Science, 31*(6), 395-418.

Manlove, S., Lazonder, A. W., & De Jong, T. (2006). Regulative support for collaborative scientific inquiry learning. *Journal of Computer Assisted Learning, 22*, 87-98.

McGrath, J. E. (1991). Time, interaction, and performance (TIP). *Small Group Research, 22*, 147-174.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning, 2*, 63-86.

Munneke, L., Andriessen, J., Kanselaar, G., & Kirschner, P. (2007). Supporting interactive argumentation: Influence of representational tools on discussing a wicked problem. *Computers in Human Behavior, 23*, 1072-1088.

Nussbaum, E. M. (2005). The effect of goal instructions and need for cognition on interactive argumentation. *Contemporary Educational Psychology, 30*, 286-313.

O'Donnell, A. M., & O'Kelly, J. (1994). Learning from peers: Beyond the rhetoric of positive results. *Educational Psychology Review, 6*, 321-349.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction, 15*, 139-178.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*, 337-386.

Saavedra, R., Earley, P., & Van Dyne, L. (1993). Complex interdependence in task-performing groups. *Journal of Applied Psychology, 78*, 61-72.

Schnotz, W., & Kürschner, C. (2008). External and internal representations in the acquisition and use of knowledge: visualization effects on mental model construction. *Instructional Science, 36*, 175-190.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences, 4*, 321-354.

Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences, 12*, 219-256.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction, 13*, 191-203.

Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education, 46*, 29-48.

Suthers, D., & Weiner, A. (1995). Groupware for developing critical discussion skills. In J. L. Schnase & E. L. Cunnius (Eds.), *Proceedings of the First International Conference on Computer Support for Collaborative Learning (CSCL'95)* (pp. 341–348). Mahwah, NJ: Lawrence Erlbaum Associates.

Suthers, D. D. (2001). Towards a systematic study of representational guidance for collaborative learning discourse. *Journal of Universal Computer Science, 7*, 254-277.

Suthers, D. D. (2003). Representational guidance for collaborative inquiry. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 27-46). Dordrecht: Kluwer Academic Publishers.

Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *Journal of the Learning Sciences, 12*, 183-218.

Teasley, S. D., & Roschelle, J. (1993). Constructing a joint problem space: The computer as a tool for sharing knowledge. In S. P. Lajoie (Ed.), *Computers as cognitive tools: Technology in education* (pp. 229-258). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). Mapping to know: The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education, 86*, 264-286.

Van Amelsvoort, M., Andriessen, J., & Kanselaar, G. (2007). Representational tools in computer-supported collaborative argumentation-based learning: How dyads work with constructed and inspected argumentative diagrams. *Journal of the Learning Sciences, 16*, 485-521.

Van Amelsvoort, M., Andriessen, J., & Kanselaar, G. (2008). How students structure and relate argumentative knowledge when learning together with diagrams. *Computers in Human Behavior, 24*, 1293-1313.

Van Bruggen, J. M., Kirschner, P. A., & Jochems, W. (2002). External representation of argumentation in CSCL and the management of cognitive load. *Learning and Instruction, 12*, 121-138.

Van Drie, J., Van Boxtel, C., Jaspers, J., & Kanselaar, G. (2005). Effects of representational guidance on domain specific reasoning in CSCL. *Computers in Human Behavior, 21*, 575-602.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner (Ed.), *Handbook of educational psychology* (pp. 841-873). New York: Simon & Schuster Macmillan.

Zhang, J. J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21*, 179-217.

Figure Captions

Figure 1: Screenshot of the Graphical Debate-tool (translated from Dutch).

Figure 2: Screenshot of the Textual Debate-tool (translated from Dutch).

Figure 3: Screenshot of VCRI.

Figure 4: Graphical representation of the topics in the available historical sources.

Table 1: Examples of coded text segments with respect to topic, grounds quality, and conceptual quality.

	Segment	Topic (see Figure 4)	Grounds quality	Concep-tual quality
1	"It is a fact that many Christians were murdered. They were used as scapegoats, and they got the blame for the economic recession. But when the Christians were executed, the economic problems remained. Therefore it was not a successful measure. Thus, it is doubtful whether they were used as propaganda for the emperor."	M2 (Christian as scapegoats)	2	1
2	"There was never a massive, full-scale persecution. The persecution was often restricted to Rome, while it was only applied in the rest of the Empire when Christians were brought before the local ruler. In the rest of the empire there has never been a systematic persecution. Also, the persecutions never lasted very long. There were always periods of peace and quiet. For example, source 13 states 'the persecution during the rule of Decius lasted for one year, at most'."	P5 (Putting the persecutions into perspective)	3	3
3	"Christians were often persecuted."	M3 (Persecution of Christians)	0	1
4	"From source 9 it becomes clear that many Christians turned themselves over to the Romans to be executed. They hoped to follow the example of Jesus Christ by sacrificing themselves. In this respect it can be concluded that many were voluntarily persecuted and this sheds a different light on the image we have of the persecutions of Christians."	P2a (Voluntary martyrs)	3	2
5	"During the Roman era, Christians were persecuted because their religion conflicted with the Roman religion. Romans were polytheistic, whereas Christians were monotheistic."	M1 (Persecution due to differences with Roman religion)	1	1

Table 2: Interrater reliability coefficients for all variables involved in the analysis of quality of essays.

	N	Agreement %	Cohen's κ	Weighted κ	Category κ range
Segmentation	4 texts	Lower bound: 81.6% Upper bound: 86.8%	–	–	–
Topic analysis	6 texts, 104 segments	87.5%	.86	–	.66 – 1.00
Grounds quality	7 texts, 113 segments	89.4%	.85	.90	.80 – .89
Conceptual quality	7 texts, 107 segments	92.2%	.88	.92	.82 – 1.00

Table 3: Collaborative activities (abbreviation in parenthesis) and category Kappas (κ_c) of the coding scheme.

	Task-related activities		*Social activities*	
	Codes	κ_c	Codes	κ_c
Performing	• Info exchange (*TaskExch*)	.89	• Greetings (*SociGree*)	.84
	• Asking questions (*TaskQues*)	.84	• Social support (*SociSupp*)	.93
			• Social resistance (*SociResi*)	.73
			• Mutual understanding (*SociUnd+*)	.95
			• Loss of mutual understanding (*SociUnd-*)	.92
Coordinating / regulating	• Planning (*MTaskPlan*)	.89	• Planning (*MSociPlan*)	.92
	• Monitoring (*MTaskMoni*)	.93	• Monitoring (*MSociMoni*)	.85
	• Positive evaluations (*MTaskEvl+*)	.89	• Positive evaluations (*MSociEvl+*)	1.00
	• Negative evaluations (*MTaskEvl-*)	.87	• Negative evaluations (*MSociEvl-*)	.67
Other	• Neutral technical (*TechNeut*)	.90	• Other / nonsense (*Other*)	.67
	• Negative technical (*TechNega*)	.90		
	• Positive technical (*TechPosi*)	1.00		

Table 4: Descriptive statistics of frequencies of adding, editing, or deleting elements in Graphical Debate and Textual Debate conditions.

	Graphical Debate		Textual Debate	
	M	*SD*	*M*	*SD*
Time online	7:43	2:03	8:26	4:35
Adding	9.82	5.82	9.80	4.32
Editing	3.53	5.43	3.07	4.85
Deleting	1.96	2.30	2.11	2.41

Table 5: Group differences for average quality of constructed representations for groups in the Graphical Debate and Textual Debate conditions.

	Graphical Debate		Textual Debate			
	M	*SD*	*M*	*SD*	*t* (37)	*d*
Number of items	15.54	6.23	18.53	5.42	-1.53	-0.50
Average quality	2.71	0.52	2.12	0.32	3.90**	1.28
- Quality Martyrs	2.75	0.57	1.98	0.54	4.17**	1.38
- Quality Propaganda	2.70	0.54	2.28	0.44	2.49*	0.83

Note. Quality was scored on a five-point scale, ranging from 0 (=low quality) to 4 (=high quality).
$^{*}p < .05$. $^{**}p < .01$.

Table 6: Mean scores and standard deviations for measures of essay quality.

	Graphical Debate		Textual Debate	
	M	*SD*	*M*	*SD*
Topics covered	6.79	1.74	7.20	2.91
Total grounds quality	29.33	7.94	25.47	9.35
Total conceptual quality	25.67	7.65	22.53	10.53
Average grounds quality	2.38	0.30	2.10	0.37
Average conceptual quality	2.06	0.22	1.82	0.42

Note. Grounds and conceptual quality were scored on a four-point scale, ranging from 0 (=low quality) to 3 (=high quality).

Table 7: Multivariate and univariate analyses of covariance for measures of essay quality.

		Univariate[b]									
	Multi-variate[a]	Topics covered		Total grounds quality		Total conceptual quality		Average grounds quality		Average conceptual quality	
	F	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
Condition (C)	3.51^{*}	2.26	.06	6.15^{**}	.15	8.30^{**}	.19	0.00	.00	7.67^{**}	.18
Length (L)	15.95^{**}	3.16	.08	3.47	.09	3.14	.08	23.53^{**}	.40	16.34^{**}	.32
C * L	3.41^{*}	1.55	.04	5.33^{*}	.13	7.70^{**}	.18	1.21	.03	3.24	.09

Note. [a] Multivariate *df* = 5, 38. [b] Multivariate *df* = 1, 39.
Condition refers to the experimental condition groups were assigned to (Graphical or Textual Debate).
Length refers to the average length of the topics written by a group in number of words.
$^{*}p < .05.$ $^{**}p < .01.$

Table 8: Descriptive statistics for pre- and post-test scores for Graphical Debate and Textual Debate conditions.

	Graphical Debate (*N* = 59)		Textual Debate (*N* = 45)	
	M	*SD*	*M*	*SD*
Pre-test	11.39	1.92	11.02	2.51
Post-test	13.08	1.76	12.22[a]	2.27

Note. Pre- and post-test performance was measured on a scale from 0 to 15.
[a] *N* = 42.

Table 9: Multilevel analyses of the effect of condition and pre-test performance on post-test performance.

	Model 0		Model 1	
	β	*SE*	*β*	*SE*
Intercept	12.78	0.27	9.60	1.07
Pre-test score			0.28**	0.10
Condition			0.42*	0.22
Variance				
Group level	1.33	0.65	0.44	0.45
Individual level	2.83	0.56	2.98	0.58
Deviance	355.70		344.63	
Decrease in deviance			11.07**	

Note. Condition was effect coded with Graphical Debate as +1 and Textual Debate as -1.
$^{*}p < .05$. $^{**}p < .01$.

Table 10: Multilevel analyses of the effect of condition on students' collaborative activities.

	β	*SE*	χ^2
Task-related activities			
TaskExch	-7.68	6.32	1.45
TaskQues	-0.76	0.75	1.01
Regulation of task-related activities			
MTaskPlan	1.11	1.73	0.40
MTaskMoni	1.84	1.62	1.25
MTaskEvl+	0.13	0.39	0.11
MTaskEvl-	-0.07	0.47	0.02
Social activities			
SociGree	0.40	0.50	0.66
SociSupp	0.13	2.83	0.00
SociResi	1.08	0.95	1.30
SociUnd+	3.12*	1.67	3.38*
SociUnd-	-1.22*	0.58	4.15*
Regulation of social activities			
MSociPlan	-0.06	1.30	0.00
MSociMoni	1.42	0.98	2.08
MSociEvl+	0.14	0.09	2.40
MSociEvl-	0.19*	0.10	3.92*
Other codes			
TechNeut	0.81*	0.48	2.78*
TechNega	-0.66	0.50	1.74
TechPosi	-0.13	0.08	2.29
Other	0.05	0.26	0.04

Note. Condition was effect coded with Graphical Debate as +1 and Textual Debate as -1.
$^*p < .05$.