MASTER'S THESIS

Stimulating Students’ Interactions and Improving the Quality of Annotations in a Social Annotation Tool by Scaffolding through Collaboration Scripts.

De Boer, Vincent

Award date:
2020

Stimulating Students' Interactions and Improving the Quality of Annotations in a Social Annotation Tool by Scaffolding through Collaboration Scripts

Het Bevorderen van Interacties tussen Studenten en het Verbeteren van de Kwaliteit van Annotaties in een Social Annotation Tool door Ondersteuning met Collaboration Scripts.

Vincent de Boer

Master Onderwijswetenschappen
Master Educational Sciences
Open Universiteit

Date: 16-10-2020
Supervisor: Dr. Howard Spoelstra

Index - Inhoudsopgave

1. Samenvatting

Achtergrond

Computer-supported collaborative learning (CSCL)-technologieën zijn ontworpen om sociaal-constructivistische leertheorieën te implementeren ten behoeve van actieve, gezamenlijke kennisconstructie, bijvoorbeeld door online discussies waarin kennis en zienswijzen worden gedeeld en beargumenteerd. De inzet van Social Annotation (SA) tools, waarbij studenten gezamenlijk annotaties schrijven bij een online academische tekst, past hier goed bij. Echter, dat studenten werken in een CSCL omgeving betekent niet dat zij automatisch deelnemen aan discussies. Eerder onderzoek suggereert dat het ondersteunen van studenten door middel van 'collaboration scripts' (instructies over samenwerking en discussie) hen stimuleert deel te nemen aan betekenisvolle interacties van hogere kwaliteit.

Doel

Het doel van deze studie is te onderzoeken hoe en in welke mate studenten, tijdens een taak in een SA tool, ondersteund moeten worden zodat zij vaker deelnemen aan argumentatieve discussies en of dit de kwaliteit van hun annotaties verhoogt.

Deelnemers, procedure, onderzoeksontwerp

In deze studie vond een experiment plaats tijdens een cursus waarin tweedejaars studenten (n=59) van een Nederlandse universiteit opdrachten uitvoerenden in de SA tool Perusall. Tijdens het experiment ontvingen de studenten in de controlegroep normale instructies, terwijl studenten in de experimentele groep zowel normale, als extra ondersteunende instructies ontvingen in de vorm van collaboration scripts.

Meetinstrumenten

Deze studie had een quasi-experimenteel ontwerp met herhaalde metingen in drie verschillende weken (nulmeting, meting na de interventie en nadat de interventie was uitgefade). In deze studie hebben we voor beide groepen gemeten welk percentage van annotaties geschreven werd als reactie op annotaties van medestudenten. De kwaliteit van annotaties werd gemeten op de niveaus van Bloom's herziene taxonomie. Om annotaties automatisch te scoren op Bloom-niveaus wilden we de 'Linguistic Inquiry and Word Count tool (LIWC2015)', gecombineerd met een lijst van (werk)woorden uit Bloom's taxonomie, valideren.

Deze validatie werd uitgevoerd door de resultaten uit een steekproef van de data te laten scoren door de tool en drie menselijke scorers en deze scores te vergelijken. Deze validatie bleek onmogelijk. Daarom werd de data, verzameld uit drie opdrachten, gescoord door de onderzoeker.

Resultaten

De resultaten uit deze studie toonden een significant verschil in percentages van annotaties, bedoeld als reactie op medestudenten, tussen beide groepen na de interventie. Echter, dit verschil werd zowel veroorzaakt door toenemende scores van de experimentele groep als door afnemende scores van de controle groep. Ook vonden we geen significant verschil binnen de experimentele groep wanneer we de scores van deze groep vergeleken over de drie verschillende meetmomenten. Voor onze kwalitatieve analyse groepeerden we de scores van annotaties op lagere en hogere cognitieve niveaus van Bloom's herziene taxonomie. Hieruit berekenden we welke percentages van annotaties binnen de hogere cognitieve niveaus vielen en vergeleken deze. Onze analyse hiervan toonde aan dat de experimentele groep significant hoger scoorde dan de controle groep na de interventie. Dit effect verdween echter nadat de extra ondersteuning was uitgefade. Tot slot vonden we geen correlaties tussen de percentages interacties tussen studenten en de percentages op de hoger cognitieve Bloom-niveaus.

Conclusie

Deze studie kon niet bevestigen dat het inzetten van ondersteuning door middel van collaboration scripts tot een significante toename van het aantal interacties tussen studenten leidde terwijl ze aan hun taak in een SA tool werkten. Wel vonden we dat studenten in de experimentele groep na de interventie hoger scoorden op de niveaus van Bloom's herziene taxonomie. Echter, dit effect verdween nadat de interventie was uitgefade.

Sleutelwoorden: Annotaties, samenwerking, Bloom

2. Summary

Background

Computer-supported collaborative learning (CSCL) technologies are used to implement social-constructivist learning theories supporting students in active and collaborative knowledge construction by encouraging students to share and discuss knowledge and arguments. The use of Social Annotation (SA) tools, in which students write annotations and engage in discussions, fits this process well. However, having students work in a CSCL environment, does not mean they automatically participate in argumentative discussions. Previous research suggests that scaffolding students’ behavior through collaborations scripts (instructions towards collaboration and discussion) encourages students to engage in more meaningful, high-quality discussions and interactions.

Aim

This study aims to examine how and to which extent students doing assignments in an SA tool need to be supported for them to engage in collaboration through online, argumentative discussions more often and whether this improves the quality of their annotations.

Participants, procedure, design

In this study an experiment took place in a second-year course of a Dutch university in weekly assignments in the SA tool Perusall (n=59). During the experiment the control group received normal instructions, while the experimental group received both normal instructions and additional scaffolding in the form of collaboration scripts.

Measures

This study had a quasi-experimental, repeated measures design, thus measurements from three different assignments (baseline, after the intervention and after the intervention had been faded out) were taken. In this study we measured the percentages of annotations that students wrote in response to fellow students. We also examined the quality of the annotations scored on the levels of Bloom’s (revised) taxonomy. Finally, we wanted to validate the Linguistic Inquiry and Word Count tool (LIWC2015) combined with a list of Bloom’s verbs for scoring annotations, by comparing the scores of the LIWC2015-tool on a sample of annotations to the scores of three human raters. We were unable to validate this instrument, meaning the annotations were scored manually by the researcher.

Results

The results of this study showed significant differences in the percentages of annotations written as a response to fellow students between the experimental and control group after the intervention. However, these differences were caused both by an increase in scores of the experimental group as well as a decrease in scores of the control group. Furthermore, we could not find significant differences within the experimental group over time in the percentages of annotations written as a response to fellow students. For our qualitative analysis we grouped the annotations of students on the lower and higher Bloom-levels of cognitive processing and calculated percentages of annotations on the higher levels. When analyzing these we found the experimental group scored significantly higher on the higher levels of cognitive processing then the control group after our intervention. This effect did not remain after the scaffolding was fully faded. We also found no correlations between student scores on percentages of interactions and percentage of annotations on the higher levels of cognitive processing.

Conclusions

This study could not confirm that the use of collaboration scripts significantly increased the number of interactions between students while working on an assignment in a SA tool. It did show students in the experimental group scored higher on the levels of Bloom’s revised taxonomy after the intervention. However, this effect did not remain over time after the scaffolding had been faded out.

Keywords: Annotations, collaboration, Bloom.

3. Introduction

3.1 Problem statement and research goal

Increasingly, computer-supported collaborative learning (CSCL) technologies have been used to implement social-constructivist learning theories in higher education and support learners in various stages of active and collaborative knowledge construction, for instance through online discussion (Gao, 2013). However, the mere fact that students are working simultaneously in CSCL environments does not automatically mean they also participate in argumentative discussions (Ertmer, Sadaf, & Ertmer, 2011; Kreijns, Kirschner, & Jochems, 2003; Valcke, De Wever, Zhu, & Deed, 2009). When the process of argumentative interaction and discussion is not properly supported, students may only focus on their own argumentation rather than contemplating the argumentation of peers. While addressing this problem, several studies on learning through online discussion boards found that students’ engagement in discussions increased through and benefited from support through scaffolding and feedback (Osborne, Byrne, Massey, & Johnston, 2018; Kobbe et al., 2007; Vogel, Wecker, Kollar, & Fischer, 2017).

Besides discussion boards, another technology for implementing CSCL in higher education is the Social Annotation (SA) tool which allows students to read academic texts online while sharing comments and questions (so-called annotations), and collaboratively prepare for lectures. Whether lectures are more focused on information delivery or more interactive and discussion oriented, proper preparation is important for students’ learning and engagement during lectures. Also, students’ understanding of literature benefits from preparing and discussing it collaboratively (Miller, Lukoff, King, & Mazur, 2018). Although reading assignments in SA tools often do encourage students to read prior to class, students (as noticed in other CSCL environments) do not necessarily engage in argumentative discussions and may only post text-related, individual annotations rather than discussing the literature’s key principles with peers (Kreijns et al., 2003; Miller et al., 2018; Valcke et al., 2009).

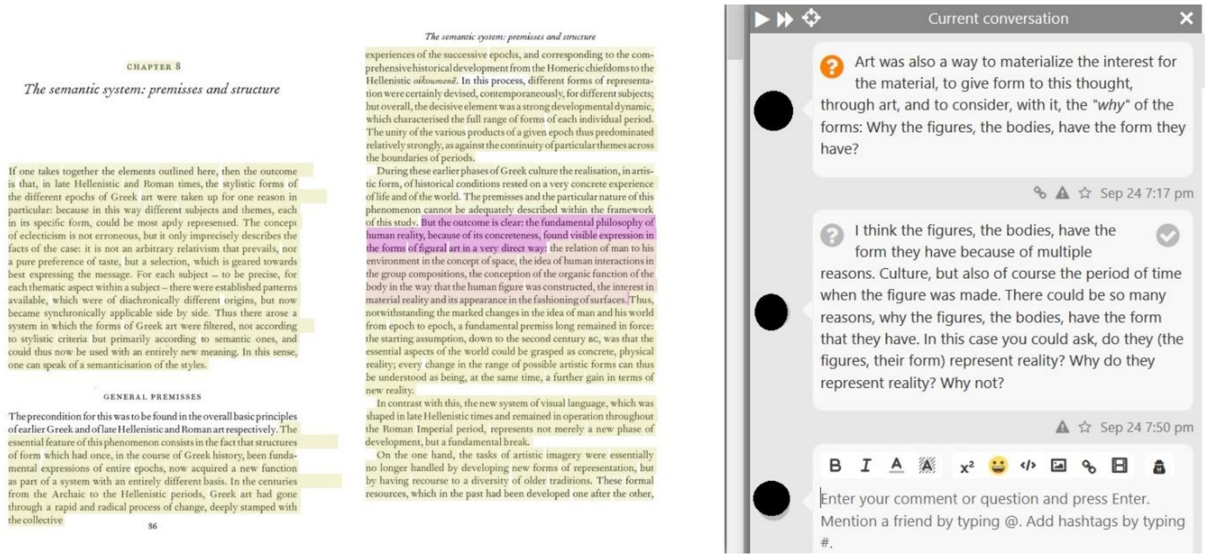
In this study we noticed that previous research on supporting collaborative learning and discussions in other CSCL environments, such as discussion boards, through scaffolding and feedback has not been widely applied to research on SA tools (Gao, 2013; Ghadirian, Salehi, & Ayub, 2018; Novak, Razzouk, & Johnson, 2012; Valcke et al., 2009). This study therefore examines how and to which extent students, using a SA tool for classroom preparation, need to be supported for them to engage in collaboration through online, argumentative discussions more often and whether this improves the quality of their annotations.

3.2 Theoretical framework

Social-constructivist learning theories, grounded in Piaget's social-cognitive and Vygotsky's socio-cultural theories, state that learning is a process of active co-construction of knowledge by individuals where knowledge is based on consensus that is constantly debated and negotiated (Duffy, 1996; Gao, 2013; Laurillard, 2009). This process of collaborative learning can be done through discussions using skills such as arguing, critical thinking and reasoning, which require learners to make their own knowledge explicit and to (re)organize their perspectives when confronted with the knowledge and ideas of others (Novak et al., 2012; Valcke et al., 2009). This is emphasized by argumentation theories used in learning sciences, showing argumentative discussion is a process of steps (such as inquiry, information-seeking and deliberation) individuals follow together to build a shared and negotiated understanding of an issue and at the same time make their own views and knowledge more explicit (Noroozi, Weinberger, Biemans, Mulder, & Chizari, 2012).

The use of ICT tools has given us new ways to engage students in collaborative learning, both in and out of the classroom. To research the effects of new, ICT driven environments for collaborative learning, researchers started using the phrase Computer Supported Collaborative Learning (CSCL) (Dillenbourg & Fischer, 2007; Schellens & Valcke, 2006). In one study on CSCL, Laurillard (2009) even stated that we should use the possibilities of implementing these CSCL-technologies to re-evaluate our education. This re-evaluation should be grounded in learning theories such as socio-cultural and collaborative learning by Piaget and Vygotsky (Laurillard, 2009) and other educational principles that CSCL connects with, such as cognitive apprenticeship, situated cognition and anchored instruction (Schellens & Valcke, 2006). When students learn through these principles they are encouraged to share and discuss knowledge and arguments. It focuses their learning and enables them to learn from peers through discussion and reflection. This process of reflection goes two ways: in (online) discussions students reflect on their peers and, being forced to contemplate their own argumentation, also reflect on their own knowledge and thought process (Laurillard, 2009; Woods & Bliss, 2016). This process was also described by Valcke et al. (2009) stating that discussions with peers force students to make their own knowledge and views explicit by retrieving their knowledge and views and putting those in the perspective of a discussion. During this process, students also negotiate the meaning of the information provided to them, because they constantly need to reorganize their own thoughts to integrate the arguments and input of others (Valcke et al., 2009). The overarching aim is that the process of collaborative learning encourages deep learning, critical thinking, shared understanding and long-term retention (Kirschner & Kreijns, 2003; Vogel et al., 2017).

The use of social annotation (SA) tools fits this process well. It offers groups of students a platform to discuss and learn about academic texts without the limitations of space and time. Students do this by performing assignments while reading academic texts online and are asked to write annotations (comment or questions about the text's content). These can both be 'new' (or 'initial') annotations with comments/ questions students have about the text, but also responses to annotations of peers, enabling (asynchronous) interaction (see Figure 1).



CHAPTER 8

The semantic system: premisses and structure

If one takes together the elements outlined here, then the outcome is that, in late Hellenistic and Roman times, the stylistic forms of the different epochs of Greek art were taken up for one reason in particular: because in this way different subjects and themes, each in its specific form, could be most aptly represented. The concept of eclecticism is not erroneous, but it only imprecisely describes the facts of the case: it is not an arbitrary relativism that prevails, nor a pure preference of taste, but a selection, which is geared towards best expressing the message. For each subject – to be precise, for each thematic aspect within a subject – there were established patterns available, which were of diachronically different origins, but now became synchronically applicable side by side. Thus there arose a system in which the forms of Greek art were filtered, not according to stylistic criteria but primarily according to semantic ones, and could thus now be used with an entirely new meaning. In this sense, one can speak of a semanticisation of the styles.

GENERAL PREMISSES

The precondition for this was to be found in the overall basic principles of earlier Greek and of late Hellenistic and Roman art respectively. The essential feature of this phenomenon consists in the fact that structures of form which had once, in the course of Greek history, been fundamental expressions of entire epochs, now acquired a new function as part of a system with an entirely different basis. In the centuries from the Archaic to the Hellenistic periods, Greek art had gone through a rapid and radical process of change, deeply stamped with the collective

86

The semantic system: premisses and structure

experiences of the successive epochs, and corresponding to the comprehensive historical development from the Homeric chiefdoms to the Hellenistic *oikoumenē*. In this process, different forms of representation were certainly devised, contemporaneously, for different subjects; but overall, the decisive element was a strong developmental dynamic, which characterised the full range of forms of each individual period. The unity of the various products of a given epoch thus predominated relatively strongly, as against the continuity of particular themes across the boundaries of periods.

During these earlier phases of Greek culture the realisation, in artistic form, of historical conditions rested on a very concrete experience of life and of the world. The premisses and the particular nature of this phenomenon cannot be adequately described within the framework of this study. But the outcome is clear: the fundamental philosophy of human reality, because of its concreteness, found visible expression in the forms of figural art in a very direct way: the relation of man to his environment in the concept of space, the idea of human interactions in the group compositions, the conception of the organic function of the body in the way that the human figure was constructed, the interest in material reality and its appearance in the fashioning of surfaces. Thus, notwithstanding the marked changes in the idea of man and his world from epoch to epoch, a fundamental premiss long remained in force: the starting assumption, down to the second century BC, was that the essential aspects of the world could be grasped as concrete, physical reality; every change in the range of possible artistic forms can thus be understood as being, at the same time, a further gain in terms of new reality.

In contrast with this, the new system of visual language, which was shaped in late Hellenistic times and remained in operation throughout the Roman Imperial period, represents not merely a new phase of development, but a fundamental break.

On the one hand, the tasks of artistic imagery were essentially no longer handled by developing new forms of representation, but by having recourse to a diversity of older traditions. These formal resources, which in the past had been developed one after the other,

Art was also a way to materialize the interest for the material, to give form to this thought, through art, and to consider, with it, the "*why*" of the forms: Why the figures, the bodies, have the form they have?

Sep 24 7:17 pm

I think the figures, the bodies, have the form they have because of multiple reasons. Culture, but also of course the period of time when the figure was made. There could be so many reasons, why the figures, the bodies, have the form that they have. In this case you could ask, do they (the figures, their form) represent reality? Why do they represent reality? Why not?

Sep 24 7:50 pm

Enter your comment or question and press Enter. Mention a friend by typing @. Add hashtags by typing #.

Figure 1. Example of an assignment and student interactions in SA tool Perusall (Screenshot), taken January 22nd 2019.

While examining the benefits of SA tools, Gao (2013) pointed out that some studies found that comments and questions of students using SA tools were richer and more focused compared to those in online discussion forums, because discussing a specific text provides students with a more focused learning activity. In a later study, while comparing SA tools to online discussion boards, Sun and Gao (2017) also found that the functionality of SA tools affords discussions to be more topic centered, instead of chronological, and contained more interrelated comments and questions, favoring the use of SA tools for online discussions over discussion boards. However, even though the use of SA tools (as a CSCL implementation) seems to fit social-constructivist learning principles very well, there is no research consensus on how and why their use improves learning (Gao, 2013). This was already noted by Valcke & Schellens (2006), but again by Gao (2013) and Ghadirian et al.'s (2018) systematic review of research on SA tools. Gao suggests that part of the problem is that most research on SA tools has been done from a variety of perspectives and models. Also, not all instruments developed for measuring learning outcomes were considered equally adequate or valid by the research community (Gao, 2013; Ghadirian et al., 2018; Valcke & Schellens, 2006).

In their study, Ghadirian et al. (2018) reviewed 71 studies on SA tools and noticed most studies focused on the implementation and evaluation of educational designs, the effects of SA tools measured through process-oriented outcomes such as metacognitive skills and critical thinking, effectiveness on learning outcomes and scores and improvement of the function of the SA tools themselves. They also found that most research did not focus on how often students actually interacted while commenting online or on what the quality of these interactions was (Ghadirian et al., 2018). However, as mentioned in our problem statement, previous research by Kreijns et al. (2003) already showed that merely placing students in (online) groups does not automatically lead to collaboration or discussion, identifying two underlying pitfalls: first, teachers often tend to focus only on the output of students towards the topic and not on the process of collaborative knowledge construction. Second, social interactions often do not spontaneously happen in collaborative environments when they are not properly supported. They suggest that scaffolding these discussions should encourage students to consider and incorporate different points of view during a collaborative learning task and to engage in more meaningful discussions and interactions (Kreijns et al., 2003). This was confirmed in 2016 by Woods and Bliss who stated that critical reflection is a skill that students do not automatically use and which should to be stimulated by teachers through scaffolding and feedback (Woods & Bliss, 2016).

Scaffolding CSCL through collaboration scripts.

The importance of scaffolding online collaborative learning was already examined by Winnips & McLoughin (2001) who pointed out that proper scaffolding stimulates learning through collaboration and discussion based on social-constructivist principles. They also suggest that the level of scaffolding should slowly be decreased (faded) during the learning process. The underlying principle of this is that scaffolding is meant to initially provide students with strategies to help them eventually self-direct their collaborative learning behavior without extra support or instruction (Winnips & McLoughin, 2001). The idea of slowly decreasing the intensity of scaffolding is supported by Vogel et al. (2017) who demonstrated that providing high intensity scaffolding for too long was experienced by students as being rigid and overly structured, taking away students' opportunities to self-direct their learning, thereby decreasing their overall motivation due to a lack of autonomy. However, they also warned that low-structured, low intensity scaffolding may not be sufficient for learners with little experience and skills in collaborative learning, and advocate that the intensity of scaffolding may need to be adapted towards learners' previous experiences (Vogel et al., 2017).

In CSCL scaffolding can be done through 'collaborations scripts' (Vogel et al., 2017). These scripts are based on Fischer's theory (2013) that collaboration skills should be regarded as internalized scripts guiding students in how to engage in the process of collaborative learning. The internalization process and development of these internal scripts can be supported by providing exemplary external collaboration scripts (instructions and examples designed to support and structure collaborative learning) encouraging learners to engage in activities such as collaboration, elaboration, explanation, argumentation and questioning (Vogel et al., 2017).

Research on collaborations scripts distinguishes between high intensity scaffolding focused on a specific learning task (for instance suggested sentence starters, question prompts or labeling to promote online discussion) called micro-scripts and low intensity scaffolding (support on the level of meta-learning, general principles and connection to the overarching learning outcomes) called macro-scripts (Kobbe et al., 2007; Prediger & Pöhler, 2015). In addition, studies by Vogel et al. (2017) and Noroozi et al. (2012) mentioned that scripts, which were specifically focused on supporting and promoting transactivity (learning through interaction with peers and responding critically to each other's contributions) amongst learners, positively affected both the collaboration between students during these learning activities itself and the domain-specific knowledge acquisition.

Noroozi et al.'s analysis on 15 years of research on Argumentations-Based CSCL (Computer Supported Collaborative Learning focusing on developing critical thinking skills based on argumentation theories) distinguished different types of scripts that can be applied, varying from explicit to implicit. These can be content-oriented scripts, facilitating the construction of declarative and procedural knowledge; and social and communication-oriented scripts, focused on the process of interaction and negotiation, to which collaboration scripts belong. They furthermore mention that the use of prompts on a social and communication level increased interaction and the level of critical thinking in discussions. (Noroozi et al., 2012). A similar distinction was made by Weinberger et al. between epistemic scripts (content/knowledge-oriented scripts) and social scripts (focused on interaction and elicitation). In the same study however, they also warned scripts should not micromanage students' activity which, especially when students gain experience in these activities, could negatively influence their sense of autonomy and motivation. Instead they should work as encouragement for them to engage in specific behavior such as collaboration and discussion (Weinberger, Ertl, Fischer, & Mandl, 2005).

Bloom's taxonomy as an instrument for measuring deep learning in CSCL and SA tools.

When it comes to reading literature, instructors often desire students to both understand and critically evaluate the texts they read. They also consider it important that students prepare for class by regularly reading course materials and actively processing information and argumentation they come across. The transfer and cognitive processing of information and argumentation in turn requires students to apply higher order cognitive skills such as comparing ideas, applying new concepts and evaluating arguments. A well-known measure for this is Bloom's taxonomy, which was originally part of three domains: cognitive, affective, and psychomotor. In education we especially focus on the cognitive domain developed by Benjamin Bloom (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). The taxonomy can also be described as levels of cognitive processing students use while working on learning activities ranging from information acquisition to more complex learning as presented by Rahman & Manaf: "1. Knowledge – the basic skill students need to obtain and remember specific pieces of information. 2. Comprehension – paraphrasing knowledge into their own words, comparing it to other knowledge and explaining a principle to others. 3. Application – students showing the ability to apply prior knowledge to new or other situations. 4. Analysis – ability to distinguish between fact and opinion and identify claims on which argumentation is build. 5. Synthesis – the need to create a new idea or product in specific situations and 6. Evaluation – critically appraise the validity of a study or knowledge and judge its relevance" (Rahman & Manaf, 2017, p. 247). These can then be bundled into lower order (knowledge, comprehension and application) and higher order (analysis, synthesis and evaluation) levels of cognitive processing, where the latter is considered a form of deep learning (Ertmer et al., 2011). Bloom's original taxonomy (Bloom et al., 1956) was later revised by Anderson & Krathwohl (2001), thereby redefining the cognitive dimension to six levels: remembering, understanding, applying, analyzing, evaluating and creating. This revised version is more in use today compared to the original and is also used in this study. Another part of the revision of Bloom's taxonomy was done by developing action verbs, making the taxonomy more applicable to describing learning outcomes and analyzing the cognitive process used by students when completing a task (Valcke et al., 2009).

When creating learning tasks in SA tools with this taxonomy in mind, most instructors do not just want students to approach academic texts through the lower levels of knowledge acquisition, comprehension and application, but also on the higher order levels of analyzing, evaluating and creating (Mulcare & Shwedel, 2017; Wang, Wei, & Ding, 2016). When analyzing students activities during learning tasks in SA tools and their interaction with peers, this taxonomy can be used to analyze the quality of online annotations of students, qualifying them from lower level annotations to the higher levels and from concrete (consisting of simple, descriptive comments and questions) to more abstract (more complex annotations showing analysis, evaluation or (co-)creation of knowledge and understanding) (Valcke et al., 2009). This is confirmed by a previous analysis of Meyer (2004) demonstrating that Bloom's taxonomy could be used for analyzing student comments in online discussions. Meyer also concluded that, although it is natural for students to engage in all levels of Bloom's thinking skills while commenting and discussing online, the percentage of higher levels comments in student's discussions can be enhanced by the type and nature of instruction given to students (Meyer, 2004). This is further supported by Ertmer et al.'s research (2011), who found that using questions or instructions constructed towards the desired outcome (of higher order thinking skills) influences the way students formulate their responses. They also found that formulating instructions that challenge students to engage in higher cognitive levels of Bloom's taxonomy led to more student-student interactions (Ertmer et al, 2011). This supports the assumption that scaffolding, by providing students with collaboration scripts, on how to engage in online collaborative discussions, should have a positive influence on both the amount of interactions students engage in and on the quality (measured on the levels of Bloom's revised taxonomy) of student's annotations while discussing academic texts in a SA tool.

3.3 Research questions and hypotheses

SA tools offer opportunities to prepare students for class by critically reading academic texts and engaging in (online) collaboration. Research suggests that providing scaffolding through collaboration scripts can increase students' interactivity while working on a learning task. This means students interact more often by responding to each other's annotations. Furthermore, using collaboration scripts focused on the process of interaction, argumentation and negotiation can increase the level of critical discussions (Noroozi et al., 2012). Research by both Meyer (2004) and Ertmer (2011) has shown that properly constructed instructions (such as collaborations scripts), focused on the desired outcome (in this case encouraging students to engage in critical discussion and argumentation), can motivate students to interact on higher cognitive Bloom-levels leading to deeper learning. Because collaboration scripts connect to higher order cognitive processing in Bloom's taxonomy, we expect an increase in the quantity and quality of the annotations of students, measured through the cognitive levels of Bloom's taxonomy, because students are not merely focused on the task of commenting on a declarative knowledge level, but are encouraged to engage in discussions leading to higher levels of cognitive processing of the texts they read. This leads to the following research questions:

RQ 1: Will students, who are scaffolded through collaboration scripts, engage in interactions/discussions more often while performing tasks on reading and annotating academic texts, in a SA environment, compared to students who do not receive scaffolding through collaboration scripts while performing the same tasks?

RQ 2: Will students, who are scaffolded through collaboration scripts, have a higher percentage of annotations counting as higher order cognitive processing on the levels of Bloom's revised taxonomy while performing tasks on reading and annotating academic texts in a SA environment, compared to students who do not receive scaffolding through collaboration scripts while performing the same tasks?

Because strong scaffolding for too long is considered as overly structured, rigid and demotivating for students and, throughout time, internalization of collaboration scripts is expected to occur in students who were supported by external collaborations scripts, scaffolds should gradually be faded to a minimum. This leads to the final research question.

RQ 3: Will the effects of scaffolding through collaboration scripts on the difference of both the quantity and quality of interactions and annotations between students who received the scaffolding, compared to students who did not receive scaffolding while performing the same task, remain over time when the scaffolding of the first group is slowly faded out during the course?

This leads to the following hypotheses:

1. Students who are scaffolded through the use of collaboration scripts, during the task of reading and annotating academic texts in an online social annotation tool, interact more often while performing their task in the SA tool, compared to students who are not scaffolded in this way.
2. Students who are scaffolded through the use of collaboration scripts more often show levels of higher order cognitive processing in their annotations (analyzing, evaluating and creating) measured on the levels of Bloom's revised taxonomy, than students who are not scaffolded in this way.
3. When the scaffolding of students, through the use of collaboration scripts, is faded out during the run time of a course, students who were initially supported through these scaffolds still show higher quantitative (number of interactions) and qualitative (higher order cognitive processing measured through Bloom's revised taxonomy) levels of interaction during the final task of reading and commenting academic texts in an online Social Annotation Tool, compared to students who were not scaffolded in this way.

The research model of this study is visualized in figure 2.

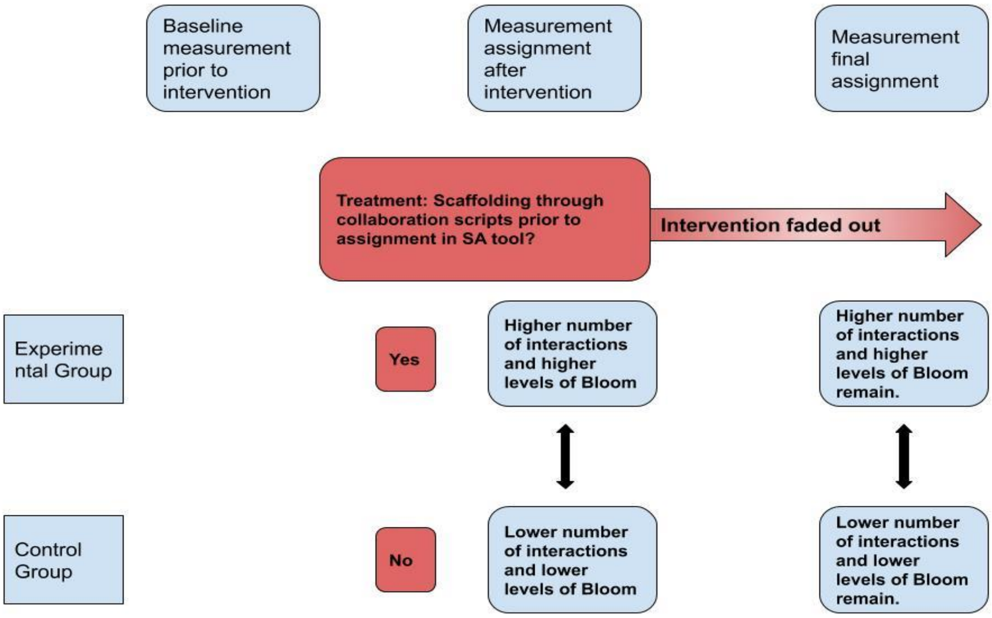

Baseline measurement prior to intervention
Measurement assignment after intervention
Measurement final assignment
Treatment: Scaffolding through collaboration scripts prior to assignment in SA tool?
Intervention faded out
Experime ntal Group
Yes
Higher number of interactions and higher levels of Bloom
Higher number of interactions and higher levels of Bloom remain.
Control Group
No
Lower number of interactions and lower levels of Bloom
Lower number of interactions and lower levels of Bloom remain.

Figure 2. Research model: faded scaffolding through collaboration scripts leads to more interactions between students and annotations of students on higher levels of Bloom's taxonomy. This effect remains over time.

4. Methodology

4.1 Design

The design of this study is a quasi-experimental, repeated measures design measuring both on a quantitative and qualitative level (Creswell, 2014). In this study an experiment was setup during reading assignments in the online SA tool Perusall over a period of seven weeks. In these assignments, students were required to read the compulsory literature for that week and place a minimum of 9 annotations per week connected to the online text. Students could decide whether the annotations would be directly related to a part of the text, or as response to an annotation of a fellow student. During this experiment students in the experimental group received scaffolding through the use of collaborations scripts next to the regular instructions for the assignments, while the control group only received regular instructions (which only explained the general assignment and requested them to read the texts and make at least 9 annotations (comments or questions) on the text in Perusall). Data was collected from the assignments of three different weeks. The scores of the experimental group and control group were compared for differences on within-group and between-group levels.

Although it was not entirely possible to exclude any chance of students sharing information about the intervention (which might affect the internal validity of the intervention through spread of the treatment or so-called selection interaction (Creswell, 2014)), the course only consisted of lectures for the entire group. Because the course only consisted of central lectures and students did the Perusall assignments individually, it was not likely students interacted about the assignments and particularly the instructions they received. Furthermore, although the students were informed about this study being conducted, they were not informed about the direction of it, since students' knowledge of the exact purpose of the study would directly have influenced the outcome of the results in both the experimental and control group, thus threatening the internal validity of this study. (Creswell, 2014). As for the external validity of this study, the students who were selected for this study are from a faculty of Arts. Although there might be slight differences in how these students interact with each other, compared to e.g., physics students, all students at the university learn and work together on academic texts and challenges through the same teaching strategy, similar to other Dutch universities. It is therefore expected the outcomes of this study are also applicable to students of other faculties and other Dutch universities.

4.2 Participants

The participants in this study were students in a second-year bachelor course at an Arts faculty from a Dutch university. In this course 102 students participated. These students were all part of the same educational program and year, hence were not randomly assigned from the entire student population of the university. The students were in the age category of 18-22. The students were randomly assigned from the entire group of students in the course to two equally sized groups. Besides taking part in the experiment, the students did not participate in other activities in these groups. The grouping mechanism in the electronic learning environment was used to show specific information to the students in these groups and send them specific instructions. For practical reasons, students were also assigned to subgroups of 10-12 students within the SA tool, to ensure the academic texts are not filled with so many annotations, that they become unreadable for students. Therefore, both the experimental and control group were split up randomly in smaller groups. Finally, it was unknown if students had previous experience in working in online annotation tools or which experience they had with collaborative learning. However, the students had been studying together up to this point in the educational program, which makes it safe to assume they roughly had the same experiences prior to this course.

4.3 Materials

For the intervention, collaboration scripts were written to scaffold the students' collaborative learning process (see Appendices A and B for the scripts used during this experiment). These scripts were based on research by Weinberger et al. (2005) and Noroozi et al. (2012) and contained instructions that aim to focus the student's attention on collaborating, making the importance of collaboration, argumentation and interaction explicit. Because students had the freedom to read the texts in their own preferred way and to come up with annotations themselves, the scripts did not structure the way and order in which students annotated the texts, providing a higher degree of freedom as recommended by Weinberger et al. (2005). Besides general instructions, Weinberger et al. (2005) and Noroozi et al. (2012) recommend using micro-scripts that offer sentence starters to start a critical discussion in any initial annotation. For instance, instead of writing 'In this segment the author connects to theory X' the students should be asked to write 'I would like to argue that the author connects to theory X, because…'. These recommendations were used to develop collaboration scripts for this study. For the experimental groups, the scripts were placed only visible to the appropriate groups in the electronic learning environment and were also emailed to them. The reason for this was that Perusall did not allow for entering additional instructions in the text on a subgroup level (so only visible for certain students), so all instructions meant as scaffolding for specific groups of students had to be placed and sent through the electronic learning environment.

For measuring the levels of higher order thinking skills, based on Bloom's (revised) taxonomy a list of verbs was compiled based on research by Bloom et al. (1956), Anderson & Krathwohl et al. (2001) and Meyer (2004). This list was used to analyze the written annotations of students for occurrences of these verbs. A non-exhaustive list of examples of these verbs can be found in table 1**.** Since the students produced over 2200 annotations, assigning them to a Bloom level manually would not be feasible during this thesis research. Therefore we explored the use of the Linguistic Inquiry and Word Count program (LIWC2015) combined with level-specific lists of Bloom-verbs to analyze the student's annotations for the occurrence of these verbs in the students' annotations. This procedure however assumed that the presence of specific verbs and words alone in individual annotations gives a proper indication of the level an annotation can be assigned to. Challenges with respect to this type of lexical or linguistic analysis were already pointed out by Kelly and Buckley (2006): first of all this type of analysis may be ambiguous. Some verbs may be present on different levels of Bloom's revised taxonomy and second, this type of analysis may be reductionist, meaning it assumes the complexity of the students' writing can be reduced to analysis through a comparison based on verbs alone. They also suggest analyzing students' discussions through more context-based analysis (Kelly & Buckley, 2006). Therefore, to validate this scoring method, a sample of 84 annotations was taken from the 3 sets of annotations available. This sample was scored by the researcher and two other educational researchers to calculate an inter-rater reliability with each other's scores and the scores of the LIWC2015-tool with the goal of validating the use of this tool for analysis.

Table 1

Examples of verbs related to the six levels distinguished in Bloom's taxonomy and Bloom's revised taxonomy (Bloom, 1956; Anderson & Krathwohl, 2001; Meyer, 2004).

Remembering	***Understanding***	***Applying***	***Analyzing***	***Evaluating***	***Creating***
Recall	*Illustrates*	*Employ*	*Dissect*	*Assess*	*Create*
State	*Characterize*	*Illustrate*	*Inspect*	*Measure*	*Improve*
Cite	*Express*	*Show*	*Correlate*	*Argue*	*Propose*
Reproduce	*Comprehend*	*Apply*	*Diagnose*	*Value*	*Synthesize*

4.4 Procedure

The experiment ran in the final quarter of the 2018-2019 academic year. The course had two teachers who were involved in the preparations of this study.

Consent

Permissions have been obtained from the teachers of the course, the educational director of the involved program and the law department of the involved institute. Instead of using a opt-in method for participation in the experiment, students were offered the opportunity to opt-out prior, during and after the experiment. The main reason for this was that past research and evaluations at the university show student responses to research requests are usually low. Besides the fact that this would have led to far less and incomplete data (strongly affecting the power of this study), it would also have made it a harder to equally distribute students over the two research groups, since this could not be done until the start of the course, because student's registrations for the course ran until that moment. This could have led to a strong unbalance between groups. Another motivation for opt-out was that the data itself was already being logged for the teachers to assess the assignments (regardless if students participated in this research or not). Both the university involved and the ethical committee of the Open University have agreed this data could be used as secondary data, in accordance to the GDPR, because the university involved has a reasonable interest to research the effectiveness of its own educational programs. Also, no extra data was requested for this study other than what was already collected and students did not encounter any negative effects from this data-processing or the intervention. Finally, students were given regular opportunities to be informed about this study and the accompanying data-analysis. Students were informed on the data processing and analysis by means of posts on and mails through the electronic learning environment. Informing students occurred in the week prior to the first lecture and during the first lecture week, when no assignment was given yet. They were given the opportunity to opt-out at any moment if they did not want their data to be examined prior to the start of the experiment, making it clear they could end their participation in this study during and after the experiment as well. For opting-out, the researchers' contact information was mailed to the students and placed on the electronic learning environment before the intervention and again when the data-processing took place.

Group distribution

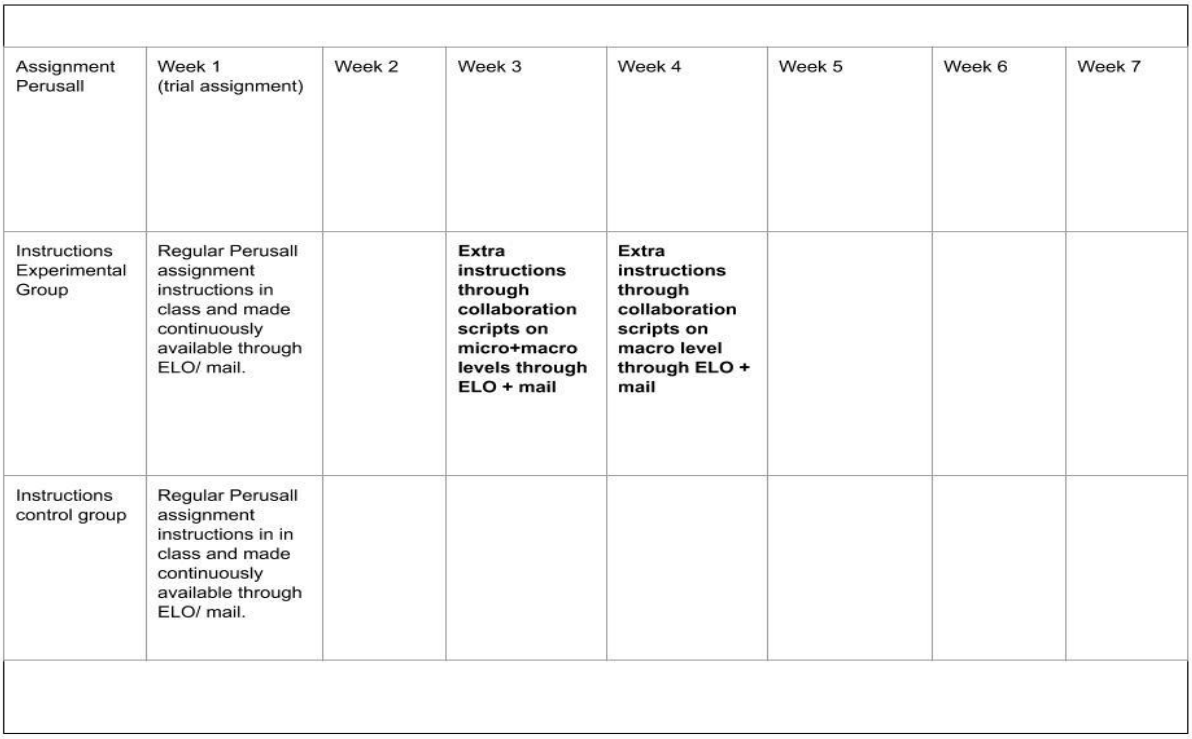
Prior to the experiment the students were distributed in 10 groups of about 10-12 students each, which is common practice for the Perusall assignments. As we had two conditions, this resulted in 5 experimental groups and 5 control groups of about equal size. All students in the experimental groups were then placed in a separate group in the electronic learning environment, so that we could provide these students with the collaboration scripts.

Experiment

All students were given seven assignments in the SA tool Perusall during the course, one for each week the course week ran. The deadlines for completion was set a few days before the corresponding lecture. Participation in the assignment of week 1 was voluntary and was not graded. It counted as a trial for all students and therefore the data it produced was not incorporated in this study. All students (from both the experimental and control groups) received the regular Perusall assignment instructions which explained how to start with the assignments. Also, as is standard for these assignments, the teachers explained in class why the Perusall assignments are done (importance of reading the academic texts and coming to class prepared). The assignment in week 2 was used for a baseline measurement, so prior to this assignment no intervention took place. For the assignment in week 3 the first intervention took place with scaffolding at its strongest, containing collaboration scripts on both a micro-level (with specific instructions consisting of suggested sentence starter) and a macro-level (with general instructions towards the process of collaborative learning). Prior to the assignment in week 4 the scaffolding was faded partially, as the students in the experimental group were only provided with collaboration scripts on a macro-level (general instructions on collaboration, but not with the specific instructions). The instructions provided the week before were still available to be reviewed in case students would prefer to look back at the instructions of week 3. For the remaining assignments the scaffolding was faded out and no extra instructions were given on collaboration during these last two assignments, although all instructions did remain available to the students. The design of the intervention is visualized in table 2.

Table 2

Design of the interventions where both groups received the regular assignment instructions through the Electronic Learning Environment (ELO), mail and in class, but only the experimental group received the additional collaborations scripts through the ELO and mail (shown in bold print).



Assignment Perusall	Week 1 (trial assignment)	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
Instructions Experimental Group	Regular Perusall assignment instructions in class and made continuously available through ELO/ mail.		**Extra instructions through collaboration scripts on micro+macro levels through ELO + mail**	**Extra instructions through collaboration scripts on macro level through ELO + mail**			
Instructions control group	Regular Perusall assignment instructions in in class and made continuously available through ELO/ mail.						

Data-collection

Our method planned on using three data sets: we started collecting data from the assignment of week 2 (with no scaffolding in place for both groups). This acted as our baseline measurement. The second data-set would be produced during the assignment of week 3 (with scaffolding on both a macro and micro level for the experimental group). The final data set would be produced by the final assignment of week 7 (when scaffolding had been faded out for several weeks). However, during the qualitative scoring process on the levels of Bloom's taxonomy, we noticed that the academic texts used in week 7 were very different in nature compared to the other texts used during the course, being more opinionated style texts, rather than the academic style of the others. This was reflected in the student's annotations, making them harder to compare to the annotations made in the previous assignments. It was therefore decided to score the annotations of week 6 (where the scaffolding for the experimental group had already been fully faded out as well) instead of week 7. For this reason we also used the data of week 6 for the quantitative analysis so that both remained comparable. Furthermore, during the experiment it became clear that students were allowed to skip one of the seven assignments, which lead to incomplete data sets produced from the different assignments for various students. Because students were allowed to skip one of their assignments, the assignments of weeks 3 and 6 (but also 7 for that matter) had a missing scores percentages in between 15-20%. Since we could not compensate for this, we only selected the output of students (n=59) who participated in all three assignments. These students were still equally divided over the experimental group (n=29) and the control group (n=30). From this student group we used 520 annotations from the assignment of week 2 (experimental group n=275, control group n=245), 536 annotations from the assignment of week 3 (experimental group n=266, control group n=270) and 539 annotations from the assignment of week 6 (experimental group n=262, control group n=277) for all our analysis.

All data was provided by the developers of the SA tool Perusall and was anonymized prior to analysis by giving each student their own ID, made up of letters and numbers, and a group-number. Each separate annotation received its own ID in the dataset. If a student responded to another student, then this annotation received an additional label matching the ID of the original annotation (so-called Parent-ID). This way we could identify whether an annotation was an initial comment or question directed at the text to read or a response to another annotation.

4.5 Analysis

In order to determine whether our scaffolding would lead to higher numbers of student interactions, our first analysis examined what percentage of students annotations were a response to an initial annotation of a fellow student versus annotations that were not a response to an annotation of a fellow student, but a stand-alone annotation directed only at the text. To score how often each student interacted with peers, each annotation received either a score of 0 (being a stand-alone annotation (not a response to a fellow student)), or a score of 1 (being a response to an annotation of a fellow student). From this, we calculated an average response score for each student per assignment which ranged from 0-1 with four decimals (so for instance 0.2091). This score can also be reported as a percentage when scaled from 0-100, which in this example would make 20,91 % of the annotations (so approximately 1 in 5) made by the student a response to fellow students and roughly 80 % of this students' annotations a stand-alone annotation (not being a response to a fellow student). These scores were on a ratio-level. The scores of the experimental group and control group were then compared. We then explored our data and tested for normality. When normality could be confirmed a Repeated Measures ANOVA, where time was the within-group variable and the intervention was the between-group variable, could be used. If not, non-parametric testing should be done by combining a Mann-

Whitney U test to compare the differences between both groups in each week and a Friedman's ANOVA-test with a split file to examine the differences over time for each group, followed up by Wilcoxon signed-rank tests with a split file to examine the difference between weeks per group. Combining these tests does mean a Bonferroni-correction for multiple testing is needed to avoid Type-I errors (Field, 2013). If non-parametric tests were needed, then for comparison we would also run and report the parametric tests (Repeated Measures ANOVA and independent t-tests (also with Bonferroni-correction)) though in that case they could not be used for our final conclusions.

For our qualitative analysis we scored the annotations of students on the levels of Bloom's revised taxonomy. As previously mentioned we wanted to use the LIWC2015-tool (combined with Bloom's verbs) for large scale analysis to determine on which level of Bloom's revised taxonomy a students' annotation would score. Both the annotations and verbs were in English. To validate the use of the LIWC2015-tool as an assessment tool for the quality of students' annotations, a sample of annotations was taken from the data and scored by three human raters who were required to assign each whole annotation to a single level of Bloom's revised taxonomy. For this, the levels of Bloom were given a value ranging from 1 point for remembering (level 1) to 6 points for creating (level 6). The scores of the human raters were compared to each other and to the results of the scores of the LIWC2015-tool. The latter was done to calculate whether the inter-rater reliability for the LIWC2015-tool could be confirmed, in which case the tool could be used for further analysis. If not, continued scoring would need to be done by the researcher. The validation of the LIWC2015-tool for assessing students' annotations was performed on a sample of 10 % (Walter, Eliasziw, & Donner, 1998) of the dataset from the first assignment. The sample consisted of 84 annotations, which were scored by three human raters. These scores were on an ordinal level. The correlation between their scores was assessed through a Spearman rank order test. If the scores of human raters amongst each other and compared to the LICW2015-tool showed a sufficiently high correlation (0.7, being a strong correlation) the inter-rater reliability of the LICW2015-tool could be confirmed (Field, 2013).

To analyze the final scores (either by the LIWC2015-tool or researcher dependent on the outcome of the validation of the tool), scores considered on Bloom's revised taxonomy in the category of lower-order cognitive processing (ranging from 1-remembering, 2-understanding and 3-applying) were grouped together and given a score of 0 and the scores considered on the level of higher-order cognitive processing (ranging from 4-analyzing, 5-evaluating, 6-creating) were given a score of 1. We then proceeded to calculate percentages of their annotations that scored on the levels of higher-order cognitive processing. These percentages were still treated on a ordinal scale because the original data was on this scale as well. This meant non-parametric testing was required to compute the differences between both the experimental and control groups over time (Field, 2013). As with the quantative analysis this was done by combining a Mann-Whitney U test to compare both groups in each week and a Friedman's ANOVA-test with a split file to examine the differences over time per group, followed up by Wilcoxon signed-rank tests with a split file to examine the difference between weeks for each group. Here too, we would perform and report parametric tests (Repeated Measures ANOVA and independent t-tests) for comparison though they could not be used for our final conclusions.

5. Results

5.1 Quantative Analysis

To analyze how often each student interacted with peers, we used the average interaction percentages calculated per student to compare the two groups. First we checked the distribution of this data and checked for outliers, first for the entire group and then for the experimental and control groups separately. A few things drew our attention: first of all, when looking at the mean scores for both groups (as can be seen in Table 3) we noticed that the percentage of annotations, being a response to fellow students, for the experimental group, increased between weeks 2 (33,5%) and 3 (44,7%) and, though decreasing between weeks 3 and 6, the mean for week 6 (39,1%) was still higher than the mean of week 2 for the this group. However, at the same time we noticed a strong decrease for the control group when comparing their means of weeks 2 (27,5%) and 3 (18%). And, though there was an increase of the mean between weeks 3 and 6, the mean of week 6 (24%) for the control group still remained lower than that of week 2. Second, we noticed that the standard deviations for all scores were very high which could be an indication our data was not normally distributed. To further examine this we looked at the frequency tables for the group as a whole and for both split groups. Here we noticed that in both groups the percentage of students who received an overall score of 0 (meaning none of their annotations were a response to a fellow student) was high: in week 2 for the control group n=8, or 26,7% and for the experimental group n=6, or 20,7%. In weeks 3 and 6 these percentages became even higher for the control group (week 3: n=13, or 43,3% and week 6: n=10, or 33,3%). For the experimental group this was however not the case: in week 3 there were no students with a mean score of 0 and in week 6 only 1 student (3,4%).

We then proceeded to calculate z-scores to look for outliers for the total group and split groups but did not see any outliers with SD=3 or higher. We therefore decided to include all available data in our analysis. Next we analyzed whether the data had a normal distribution and again noticed a high distribution and, as shown in Table 3**,** all data showing high skewness. We applied the Kolmogorov-Smirnov test for normality and found the distributions for the total group of students on percentages of interactions significantly non-normal for week 2, *D*(59)= 0.131, *p*= .013 and for week 6, *D*(59)= 0.132, *p*= .012. For week 3 no significant deviation from normal was found, D(59)= 0.114, *p*= .051.

The split scores per group were mostly significantly non-normal too.

Table 3

Means, Medians, Standard Deviations and Standard Errors for the total group, the control group and experimental group on levels of student interactions per assignment.

Assignment	**Group**	**Annotations**	**M**	**Mdn**	**SD**	**SE**	**Skewness**
Week 2	Total (N=59)	N=520	.305	.222	.264	.034	.800
	Control (N=30)	N=245	.275	.222	.270	.049	1.189
	Experimental (N=29)	N=275	.335	.316	.258	.048	.457
Week 3	Total (N=59)	N=536	.311	.333	.258	.034	.700
	Control (N=30)	N=270	.180	.095	.233	.042	1.713
	Experimental (N=29)	N=266	.447	.444	.210	.039	.844
Week 6	Total (N=59	N=539	.314	.333	.251	.033	.954
	Control (N=30)	N=277	.240	.222	.257	.047	1.318
.	Experimental (N=29)	N=262	.391	.363	.222	.041	1.189

We therefore decided non-parametric testing needed to be conducted, combining the Mann-Whitney U test and a Friedman's ANOVA-test with a split file, followed up by Wilcoxon signed-rank tests with a split file to examine the difference between weeks per group. Because our analysis combined multiple test (7 tests in total) we used a Bonferroni-correction for our non-parametric tests with α (0.05/7)= .0071 as significance level. Though these non-parametric tests results formed our main analysis, for comparison we also performed parametric tests.

First Mann-Whitney U tests were performed to check for differences between groups per week. To make a comparison to the Mann-Whitney U tests, we ran independent, two-tailed t-tests for the same weeks for which we also corrected for multiple testing over three tests (α (0.05/3)= .0167) and analyzed the Repeated Measures interaction effects. For week 2 the Mann-Whitney U test (as can be seen in table 4) showed that the median of percentages of annotations being a response to fellow students of the control group did not differ significantly from that of the experimental group. The same result can be seen in the parametric t-test for week 2. Being the baseline measurement, this confirmed there were no significant differences between both groups prior to the intervention.

Table 4

Comparing the results between groups with both non-parametric testing and parametric testing on percentages of interactions by students.

Mann-Whitney U test	**Control group *Mdn* *(*n=30)**	**Experimental group *Mdn* (n=29)**	**Result (α = .0071)**	**Effect size Pearson's *r***
Week 2	.222	.316	*U*= 504, *z*= 1.055, *p*= .291	*r* =.13
Week 3	.095	.444	*U*= 721, *z*= 4.375, *p*< .001	*r* =.57
Week 6	.222	.363	*U*= 620.5, *z*= 2.832, *p*= .005	*r* = .37
Independent t-test	**Control group mean *(*n=30)**	**Experimental group mean (n=29)**	**Result (α= .0167)**	**Effect size Cohen's *d*.**
Week 2	.275	.335	*t*(57)= .863, *p*= .392	*d*= *.23*
Week 3	.180	.447	*t*(57)= 4.631, *p* <.001	*d*= 1.21
Week 6	.240	.391	*t*(57)= 2.450, *p*= .019	*d*= .63

However, the median in week 3 of the experimental group was significantly higher than that of the control group with *U*= 721, *z*= 4.375, *p*< .001, *r* =.57 being a large effect. For week 6 the median of the experimental group was also significantly higher than that of the control group with *U*= 620.5, *z*= 2.832, *p*= .005, *r* = .37, being a medium effect. The independent t-test for week 3 is in line with the non-parametric test showing *t*(57)= 4.631, *p* <.001, *d* =1.21 also being a significant difference with a large effect. For week 6 however, due to a stricter alpha, the t-test did not confirm the non-parametric test. We also compared these results to the interaction effect of the Repeated Measures ANOVA test (table 5). For analysis of this test we first checked an equal distribution in the variance through Mauchly's test of sphericity. This showed the assumption of sphericity was met, $\chi2(2) = 5.640$, *p* = .06. The Repeated Measures ANOVA showed a significant interaction-effect between the percentages of students' annotations being a response to fellow students, and the (experimental or control) group those students were in, *F* (2,114) = 6.922, *p* =.001, *r* = .06, being a very small effect. However, when looking at the simple contrasts comparing weeks 3 and 6 to week 2, we saw this effect was only significant for the contrast between weeks 2 and 3, *F* (2,57) = 10.632, *p* = 002, *r*= .16 (being a small effect), but not significant for the contrasts between weeks 2 and 6, *F* (2,57) = 3.432, *p* = .069, *r* = .06.

Though overall our results showed significant differences between the experimental group and control group after the intervention, we want to emphasize these can be attributed to both the increase in scores of the experimental group as well as the decrease in scores of the control group, especially in week 3.

Table 5

Comparing the results of statistical tests on percentages of interactions by students through non-parametric and parametric tests.

Non-parametric and parametric ANOVA’s			
Friedman’s ANOVA with split file per group (α = .0071)	Control group (n=30): *χ2*(2) = 6.871, *p* = .032	Experimental group (n=29): *χ2*(2) = 2.440, *p* =.295	
Repeated Measures ANOVA simple effects with split file per group (α = .01)	Control group (n=30): *F* (2,58) = 3.142, *p* =.051, *f* = .27	Experimental group (n=29): *F* (2,56) = 3.814 *p* =.028, *f* = .31	
Repeated Measures ANOVA Interaction effect _time*group_ (α = .05)	*F* (2,114) = 6.922, *p* =.001, *r* = .06		
Post hoc tests			
Wilcoxon test	**Difference week 2-3 (α = .0071)**	**Difference week 2-6 (α = .0071)**	**Difference week 3-6 (α = .0071)**
Control group (n=30)	*T* = 53, *p*= .031, *r*= .0,28	*T* = 90, *p*= .373, *r*= .012	*T* = 131.5, *p*= .142, *r*= .19
Experimental group (n=29)	*T* = 311.5 , *p*= .042, *r*= .027	*T* = 207, *p*= .231, *r*= .016	*T* = 120, *p*= .159, *r*= .18
Post Hoc (pairwise comparison) Repeated Measures ANOVA with split file	**Mean difference week 2-3 (α= .01)**	**Mean difference week 2-6 (α= .01)**	**Mean difference week 3-6 (α= .01)**
Control group (n=30)	-.096, *p*= .137	-.035, *p*= .900	.061, *p*= .297
Experimental group (n=29)	.112, *p*= .051	.056, *p*= .401	-.056, *p*= .543

We then preformed a Friedman’s ANOVA test with split files for the experimental and control groups, and Wilcoxon tests as post-hoc tests to compare the differences between weeks within the groups (α = .0071). For comparison, we computed simple effects for each group and performed post-hoc tests (using pairwise comparisons) by repeating the Repeated Measures ANOVA with split files per group with a Bonferroni-correction for multiple testing over 5 tests (2 for the main effects and 3 post hoc tests, α = .01). These results can also be seen in Table 5. The Friedman's ANOVA test showed that the percentages of annotations being a response to fellow students for both the control group and experimental group did not significantly change throughout the three measurements over time. This is confirmed by the simple effects from the Repeated Measures ANOVA, showing no significant changes over time for each group. The follow-up non-parametric Wilcoxon signed-rank tests showed no significant differences for both the control and experimental groups between weeks 2 and 3, weeks 2 and 6 and weeks 3 and 6. The post-hoc tests from the Repeated Measures ANOVA using the pairwise comparisons confirm these results. As with the non-parametric testing none of the compared means were statistically different for either group when comparing the results from weeks 2, 3 and 6 meaning no significant differences in scores within the groups over time can be confirmed. Therefore we establish these tests showed that no significant differences between the scores throughout the weeks for the both the control and experimental groups could be confirmed.

5.2 Qualitative Analysis

We aimed to use the LIWC2015 linguistic analysis tool combined with verbs from Bloom's (revised) taxonomy for qualitative analysis. To validate the results from the tool, the assignments to Bloom levels of a sample of the data performed by three human raters and the LIWC2015 tool were compared. First, the researcher and two other human raters scored a selection of 10 annotations from the assignment of week 2 and had a calibration session. For scoring the human raters received instructions (see Appendix D). They also read the articles the students commented on first, so they would have a better understanding of the topic students were referring to. During the calibration session we found that a few decisions for further scoring needed to be made: 1. The level from the Bloom taxonomy scored should fit the definition of that level. Sometimes a rater doubted if a comment could be scored on a particular level and then decided to score one level lower on the scale without checking whether the definition of this level actually matched with the annotation. 2. In many cases we noticed students attempted to write annotations fitting a certain level of the taxonomy, but did a poor job performing on this level. Given that the assignments were situated in the course prior to the lectures and that students had not encountered the material before, we decided to score the intention of the student regardless if they succeeded or not. As this study examined whether instructions towards collaborative learning can influence student behavior we believed scoring students' intended behavior to be plausible.

Table 6

Correlations between the human raters and the LIWC2015 scores calculated with the Spearman rank-order correlation coefficient.

Score n=84	Scorer 1	Scorer 2	Scorer 3	LIWC mean All levels	LIWC mean Highest level	LIWC Median	LIWC Prevelant	LIWC Highscore
Scorer 1	X	r_s=.471, *p<.001*	r_s=.467, *p<.001*	r_s=.124, *p=.26*	r_s=.134, *p=.23*	r_s=.192,*p =.08*	r_s=.072,*p =.51*	r_s=.038,*p=. 73*
Scorer 2		X	r_s=.747,*p <.001*	r_s=.126, *p=.15*	r_s=.160, *p=.15*	r_s=.159, *p=.15*	r_s=.144, *p=.19*	r_s=.04, *p=.72*
Scorer 3			X	r_s=.155, *p=.16*	r_s=.148, *p=.18*	r_s=.193, *p=.08*	r_s=.,099 *p=.37*	r_s=.046, *p=.68*
LIWCmeanalllevels				X	r_s=.955, *p=<.001*	r_s=.927, *p=<.001*	r_s=.832, *p=<.001*	r_s=.674, *p=<.001*
LIWCmeanhighestlevel					X	r_s=.897, *p=<.001*	r_s=.773, *p=<.001*	r_s=.653, *p=<.001*
LIWCmedian						X	r_s=.780, *p=<.001*	r_s=.620, *p=<.001*
LIWCprevelant							X	r_s=.572, *p=<.001*
LIWChighscore								X

As the LIWC2015-tool searched for the presence of words from the predefined lists of verbs/words from Bloom's taxonomy (see Appendix C) in annotations, it could give each annotation scores on various levels (because words or verbs from multiple levels can be present in one annotation). Therefore several choices needed to be made: 1) Whether to use the mean or the median score coming from the tool? We therefore calculated both and compared both to the scores of the human raters. We also looked at the most prevalent level in each annotation (most counts) and calculated scores when only taking the highest score into account. 2) As some verbs appeared on multiple Bloom levels and previous research shows that some verbs can match different levels of the taxonomy, we calculated scores for both, meaning: a. allowing LIWC2015 to score a verb/word on several levels and b. allowing LIWC2015 to only score the highest level. Since the variables were on an ordinal level, the correlations were calculated using the Spearman's rho. As can be seen in Table 6 there were significant and moderate to strong correlations between all three human raters, where the correlation between scorer 1 and scorer 2 was *rs*(84)=.471, *p*<.001, the correlation between scorer 1 and scorer 3 was *rs*(84)=.467, *p*<.001 and the correlation between scorer 2 and scorer 3 was *rs*(84)=.747, *p*<.001. Furthermore, the test also showed high, significant correlations between the different interpretations of the LIWC2015 scores. There were no significant correlations between any of the scores of the LIWC2015-tool and the human raters. Based on these results, using the results from LIWC2015-tool (combined with the verb/words list from Bloom's (revised) taxonomy, as an analysis tool for the levels of Bloom that students' annotations are on, could not be validated.

Therefore, the final scoring was performed by the researcher on the annotations of the three weeks from the data-set. To avoid bias, the researcher scored the annotations in an excel-file without reference to what group each student was in. Then the scores were reconnected the annotations and the group they belonged to. After scoring the annotations, the scores considered in the category of lower-order cognitive processing (ranging from 1-remembering, 2-understanding and 3-applying) were given a score of 0 and the scores considered on the level of higher-order cognitive processing (ranging from 4-analyzing, 5-evaluating, 6-creating) were given a score of 1. We then calculated percentages of their annotations that were scored in the category of higher-order cognitive processing. This meant that if a student had a score of 0.2091, 20,91% of this student's scores fell into the category of higher-order cognitive processing.

We however also created a table of the scores from both groups for all six levels of Bloom's revised taxonomy in order to gain insight in how each group of students had scored on all levels per week. These scores, as shown in table 7, showed both groups initially scored relatively high on the level of remembering with 37.1% of the scores of the experimental group and 40.8% of the scores of the control group and showed the scores on the levels of higher order cognitive processing being relatively lower. This pattern seemed to continue for the control group throughout the weeks, but the scores of the experimental group showed a different picture after week 3. Though the largest percentage of scores still fell within the lower order cognitive processing levels, we did see an increase of (percentages) of scores on the higher order cognitive processing levels, especially on the scores of 'Evaluating', from n=33 (12%) to n=70 (26.3%). Scores on this level were given when students not only agreed or disagreed with the author or fellow students, but explicitly tried to support their claims with ideas, theories and knowledge. We also noticed after the intervention in week 3 the percentage of annotations on the levels of 'Remembering' dropped for the experimental group, where it did not (or even increased) for the control group. Looking at the scores of week 6, for the experimental group, the scores on lower order cognitive processing levels seemed to increase again, but we did notice that the scores on the level of 'Evaluating' remained relatively high and for 'Remembering' relatively low.

Table 7

Table of Bloom-scores for weeks 2, 3 and 6 for the experimental and control group.

	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
Experimental group						
Week 2 (n=275)	102 (37.1%)	76 (27.6%)	42 (15.3%)	22 (8%)	33 (12%)	-
Week 3 (n=266)	67 (25.2%)	75 (28.2%)	32 (12%)	21 (7.9%)	70 (26.3%)	1 (0.4%)
Week 6 (n=262)	52 (19.8%)	105 (40.1%)	27 (10,3%)	22 (8.4%)	56 (21.4%)	-
Control group						
Week 2 (n=245)	100 (40.8%)	64 (26.1%)	35 (14.3%)	14 (5.7%)	32 (13.1%)	-
Week 3 (n=270)	135 (50%)	68 (25.2%)	26 (9.6%)	13 (4.8%)	27 (10%)	1 (0.4%)
Week 6 (n=277)	99 (35.7%)	98 (35.4%)	38 (13.7%)	11 (4%)	31 (11.2%)	-

Finally, we noticed both group-scores were consistently high on the level of understanding, with an increase for the both groups in week 6. The relatively consistent scoring on this level is not surprising given the nature of the Perusall assignments and the fact that these texts were used to prepare students for class and that this material was relatively new to them. The scores on week 6 may indicate that students struggled more to understand the texts of this week compared to others.

We then proceeded to explore the data's means, medians, standard deviations, standard errors and skewness as can be seen in table 8. As in our quantitative analysis, we noticed a high standard deviation around the mean and in several cases a high skewness. Here too we noticed that the scores for the experimental group increased between weeks 2 (20.1%) and 3 (31.4%) and, though decreasing between weeks 3 and 6, the mean for week 6 (23.7%) also remained higher than the mean of week 2. Though we did see a decrease in the mean of the control group when comparing their means of weeks 2 (19,5%) and 3 (15.6%), the decrease (3.9%) was less severe than in our quantitative analysis (which was 9.5%) and the scores stayed more consistent over time.

Table 8

Means, Medians, Standard Deviations and Standard Errors for percentages of student annotation on the level of higher-order cognitive processing from Bloom’s revised taxonomy.

Assignment	**Group**	**Annotations**	**M**	**Mdn**	**SD**	**SE**	**Skewness**
Week 2	Total (N=59)	N=520	.198	.167	.178	.023	1.252
	Control (N=30)	N=245	.195	.163	.156	.028	.479
	Experimental (N=29)	N=275	.201	.167	.200	.037	1.622
Week 3	Total (N=59)	N=536	.233	.222	.173	.023	.662
	Control (N=30)	N=270	.156	.118	.122	.022	.421
	Experimental (N=29)	N=266	.314	.333	.183	.034	.289
Week 6	Total (N=59)	N=539	.226	.222	.176	.023	.179
	Control (N=30)	N=277	.215	.222	.180	.033	.096
.	Experimental (N=29)	N=262	.237	.222	.175	.033	.297

To further explore our data we looked at both the frequency tables for the group as a whole and for both split groups. In the frequency table we noticed that in week 2, for both groups, the number and percentages of students who received an overall score of 0 (meaning all of their scores scored on the levels of lower order cognitive processing) were relatively high (control group n=6, or 20% and for the experimental group n=7, or 24.1%). In weeks 3 and 6 these percentages became even higher for the control group (week 3: n=7, or 23,3% and week 6: n=9, or 30%). For the experimental group this was however not the case: in week 3 there were 3 students (10.3%) with an overall score of 0 and in week 6 there were 5 students (17.2%). We then proceeded to calculate z-scores to look for outliers. When looking at the z-scores for the total group and all weeks, we spotted only 1 outlier with SD=3 or higher in the total group (SD = 3.89, belonging to the experimental group) in week 2. We therefore decided we could include all data in our analysis. Next an analysis of the data was done to see whether the data had a normal distribution. We applied the Kolmogorov-Smirnov test for normality and found the distributions for the total group of students on percentages of levels of higher order cognitive processing were significantly non-normal for week 2, *D*(59)= 0.151, *p*= .002 and for week 6, *D*(59)= 0.137, *p*= .008. For week 3 no significant deviation from normal was found.

Being that the data was also on an ordinal scale, non-parametric testing was required, combining the Mann-Whitney U test and the Friedman’s ANOVA-test with a split file, followed up by

Wilcoxon signed-rank tests with a split file to examine the difference between weeks per group. Because we used multiple testing (7 tests combined) for our analysis, we used a Bonferroni correction for multiple testing to prevent Type I-errors, were we used α (0.05)/7= .0071 as significance level. Though the non-parametric tests became our main analysis, we compared these results to parametric tests. First Mann-Whitney U tests were performed to compare the scores of both groups per week (Table 9). To make a comparison to the Mann Whitney U test, we ran independent, two-tailed t-tests for the same weeks. Here too we corrected for multiple testing over three tests (α (0.05)/3= .0167).

Table 9

Comparing the results between groups through both non-parametric testing (Mann Whitney U tests) and parametric testing (independent t-tests) for measuring higher order cognitive processing.

Mann Whitney U test	Control group *Mdn (*n=30)	Experimental group *Mdn* (n=29)	Results (α = .0071)	Effect size Pearson's *r*
Week 2	.163	.167	*U*= 421, *z*= -.214, *p*= .831	*r* =-.27
Week 3	.118	.333	*U*= 671, *z*= 3.593, *p*< .001	*r* =.47
Week 6	.222	.222	*U*= 468.50, *z*= .512, *p*= .609	*r* = .067
Independent t-tests	Control group mean *(*n=30)	Experimental group mean (n=29)	Results (α= .0167)	Effect size Cohen's *d*.
Week 2	.195	.201	*t*(57)= -.127, *p*=.900	*d*= .03
Week 3	.156	.314	*t*(57)= 3.904, *p* <.001	*d*= 1.01
Week 6	.215	.237	*t*(57)= .471, *p*= .639	*d*= .12

The median of percentages of annotations being scored on the levels of higher order cognitive processing in week 2 and week 6 of the control group did not differ significantly from that of the experimental group. With week 2 being the baseline measurement, this confirmed there were no significant differences between both groups prior to the intervention. The same result was seen in the independent t-test of week 2 showing no significant difference. However, the median score in week 3 of the experimental group (*Mdn*= .333) was significantly higher than that of the control group (*Mdn*= .118), *U*= 671, *z*= 3.593, *p*< .001, *r* =.47 being a medium effect. This too can be seen in the independent t-test showing *t*(57)= 3.904, *p* <.001, d = 1.01, being a significant difference between both groups with a large effect. We also compared these results to the interaction effects of the Repeated Measures ANOVA test (table 10).

Table 10

Comparing the results of statistical tests on percentages of Bloom-levels of higher order cognitive processing by students through non-parametric testing and parametric testing.

Non-parametric and parametric ANOVA's			
Friedman's ANOVA with split file per group (α = .0071)	Control group (n=30): *χ2*(2) = 2.262, *p* = .323	Experimental group (n=29): *χ2*(2) = 6.288, *p* = .043	
Repeated Measures ANOVA simple effects with split file per group (α = .01)	Control group (n=30): *F* (2,1.360) = 1.159, *p* =.306, *f* = .06	Experimental group (n=29): *F* (2,1.783) = 3.227 *p* =.053, *f* = .05	
Repeated Measures ANOVA Interaction effect *time*group* (α = .05)	*F* (2,90.3) = 3.879, *p* =.033, *r* = .04		
Post hoc tests			
Wilcoxon test (non-parametric)	**Difference week 2-3 (α = .0071)**	**Difference week 2-6 (α = .0071)**	**Difference week 3-6 (α = .0071)**
Control group (n=30)	*T* = 84.5, *p*= .103, *r*= -.212	*T* = 205.5, *p*= .692, *r*= .052	*T* = 275, *p*= .101, *r*= .21
Experimental group (n=29)	*T* = 307.5 , *p*= .004, *r*= .0,37	*T* = 248, *p*= .305, *r*= .013	*T* = 107.5, *p*= .05, *r*= -.26
Post-hoc test (pairwise comparison) from Repeated Measures ANOVA with split file	**Mean difference week 2-3 (α= .01)**	**Mean difference week 2-6 (α= .01)**	**Mean difference week 3-6 (α= .01)**
Control group (n=30)	-.040, *p*= .365	020, *p*= 1.00	.059, *p*= .421
Experimental group (n=29)	.112, *p*= .011	035, *p*= 1.00	-.077, *p*= .311

For proper analysis of the Repeated Measures ANOVA we first checked an equal distribution in the variance through Mauchly's test of sphericity. This showed the assumption of sphericity was not met, $\chi2(2) = 21.005$, p <.001. Since the estimate of the Greenhouse-Geisser correction was greater than the acceptable limit of .75, being .762, we used the Huynh-Feldt corrected output. This test showed a significant statistical interaction-effect between the percentage of each students' annotations being scored on the levels of higher order cognitive processing, and the (experimental or control) group those students were in, *F* (2,90.3) = 3.879, *p* =.033, *r* = .04, being a very small effect. When looking at the simple contrasts comparing weeks 3 and 6 to week 2, we saw this effect was only significant for the contrast between weeks 2 and 3, *F* (2,57) = 12.447, *p* = 001, *r*= .18 (being a small effect), but not significant for the contrasts between weeks 2 and 6, *F* (2,57) = .047, *p* = .829.

We then preformed the Friedman's ANOVA test, with split files for the experimental and control groups, and Wilcoxon tests as post-hoc tests to compare the differences between weeks within both group (α = .0071). For our comparison, we computed simple effects for each group and performed post-hoc tests (using pairwise comparisons) by repeating the Repeated Measures ANOVA with split files per group with a Bonferroni-correction for multiple testing over 5 tests (2 for the main effects and 3 post hoc tests, α = .01). These results can also be seen in Table 10. The Friedman's ANOVA test showed that the percentages of annotations being scored on the levels of higher order cognitive processing for both the control group and experimental group (the latter partially due to the stricter α) did not significantly change throughout the three measurements over time. This is confirmed by the simple effects from the Repeated Measures ANOVA, showing no significant changes over time for each group. The follow-up Wilcoxon signed-rank tests for the experimental group did show a significant difference between weeks 2 and 3, *T* = 307.5 , *p*= .004, *r*= .0,37, being a medium effect. We saw no significant differences for the experimental group between weeks 2 and 6 and between weeks 3 and 6. For the control group, the Wilcoxon signed-rank tests showed no significant differences between all weeks. The post-hoc tests from the Repeated Measures ANOVA with split file, using pairwise comparisons, confirmed most of these results, but not the significant difference between weeks 2 and 3 for the experimental group, which was also non-significant due to a stricter α.

Comparing the quantitative and qualitative scores

Finally, we were interested to see if students who scored a higher percentage of interactions also scored a higher percentage of scores on higher order cognitive processing levels of Bloom's revised taxonomy. Being that for both scores normal distributions could not be confirmed we used the non-parametric Spearman's Rho test for correlation to compare these results. When looking at the results, as shown in table 11, we only found a significant, low correlation when looking at the scores of all students, between the scores on percentages of interactions and percentages of annotations on the higher order levels of Bloom, for week 3. When splitting the groups however, this correlation did not appear for week 3 in either group. For all other weeks no significant correlations were found between the scores. We can therefore not confirm any correlation exists between student scores on the percentages of interactions and their scores on percentages of annotations scored on the levels of higher order cognitive processing from Bloom's revised taxonomy.

Table 11

Correlations (R_s) between the scores on percentages of interactions between students and percentages of annotations scores on levels of higher order cognitive processing from Bloom's revised taxonomy.

Total group	**Week2 Bloom**	**Week3 Bloom**	**Week6 Bloom**
Week2 Interactions	r_s= -.007, *p= .958*		
Week3 Interactions		r_s= .373, *p= .004*	
Week6 Interactions			r_s= .201, *p= .128*
Control group	**Week2 Bloom**	**Week3 Bloom**	**Week6 Bloom**
Week2 Interactions	r_s= -.121, *p= .525*		
Week3 Interactions		r_s= .127, *p= .503*	
Week6 Interactions			r_s= .133, *p= .482*
Experimental group	**Week2 Bloom**	**Week3 Bloom**	**Week6 Bloom**
Week2 Interactions	r_s= .101, *p= .602*		
Week3 Interactions		r_s= .224, *p= .244*	
Week6 Interactions			r_s= .171, *p= .376*

6. Discussion, conclusion, limitations and future research

6.1 Discussion

As far as the main findings of this study go, we could not establish that the experimental group's scaffolding, based on collaborations scripts, led to a significant increase in interactions between students in this group. Though the experimental group did score significantly better than the control group in weeks 3 and 6, this could be contributed to both an increase in scores of the experimental group as a decrease in scores of the control group. Furthermore, the descriptive statistics showed a high distribution around the mean and a strong skewness to the lower scores regardless of group. This showed that, even after the intervention, a large group of students in the experimental group still scored low on the interaction percentages. One possible explanation could be whether the task in the SA tool itself contributed to the willingness or motivation to collaborate amongst the students? As Kirschner, Paas & Kirschner (2011) showed, collaborating with others on a task requires an investment in time and effort. The benefits of collaborating thus need to be worth a students' time and effort. This means a task needs to be complex and/or challenging enough to make collaborating more beneficial then doing the task on your own. It is possible that the Perusall tasks, or the way these tasks were presented, were found to be repetitive or not challenging enough to make collaboration feel worthwhile for part of the students regardless of the scaffolding.

Though our study did find a significant change in the scores on the percentage of annotations being on the levels of higher order cognitive processing in week 3 for the experimental group, this effect did not remain after the scaffolding was fully faded. Besides the question whether the task itself made collaborating worthwhile, another possible explanation for this may be that the course's teachers were not asked to follow-up the use of collaboration scripts with feedback on how students had engaged in collaborative argumentation during the assignments and only provided the entire group of students feedback on annotations related to the content of the academic texts itself. Furthermore, the design of this study set out to measure the effects of the instructions and not a combination of instruction and feedback. It is however possible that a combination of instructions and feedback would have stimulated the students in the experimental group to stay engaged in collaborative behavior over a longer period of time. This also connects to work by Bloom (1956) and Hattie (2012) which showed that a combination of different instructional methods often leads to greater success compared to only implementing one intervention or method.

Regarding our attempt to validate the use of the LIWC2015 tool for qualitative analysis in our study, this study confirms previous work of Kelly and Buckley (2006) stating that linguistic analysis with verbs from Bloom's Taxonomy may be ambiguous, because some words are present on different levels of Bloom's taxonomy and reductionist, assuming the complexity of the students' writing can be reduced to analysis based on a linguistic comparison to single words. During the calibration sessions the human raters also noticed the ambiguity of the taxonomy as a whole, meaning they had to decide whether to 'grade' based on the actual quality the student annotation or the level of the outcome we perceived the student attempted to reach. Given the fact that the Perusall assignments took place prior to classes, it seemed most logical to score the intention of the student. The use of the LIWC2015-tool for analysis with Bloom-verbs posed similar issues: because the LIWC2015-tool could score different Bloom-levels within a single annotation, we had to choose how to use and interpret these scores. This meant there was a risk of choosing an interpretation of the data of the LIWC2015-tool which did not match the approach the human raters had taken. To solve this problem, we decided to compare the scores of the human raters with several interpretations of the LIWC2015-scores. Even with this method however, we found no correlations between the scores from the LIWC2015-tool and the scores of the human raters and were unable to validate the results of the LIWC2015-tool combined with Bloom verbs. A reason for this could be that the human raters had a calibration session during the scoring process to agree on the choices that were made for scoring the students' annotations. They also read the academic texts the students had used to write annotations on. This means their scores were contextualized unlike to the scores of the LIWC2015-tool which made simpler comparisons to verbs.

6.2 Conclusion

Our first research question was whether students, who are scaffolded through collaboration scripts, engage in interactions/discussions more often while performing tasks on reading and annotating academic texts, in a SA environment, compared to students who do not receive scaffolding through collaboration scripts while performing the same tasks? During our analysis we saw an increase in percentages of annotations being a response to fellow students in the experimental group after the intervention in week 3. Though this percentage decreased in week 6, it still remained higher than the percentage in week 2. We also saw the opposite (a relatively sharp decrease of a 9,5 %) in the percentages of annotations being a response to fellow students for the control group in week 3. The Mann-Whitney U tests did show significant differences in scores after the intervention and after the fading of the intervention, when comparing the scores of both groups. However, these differences could be attributed to both the increase in the mean score of the experimental group, as the decrease in the mean score of the control group. Finally, looking at the results of the Wilcoxon signed-rank tests, we could not confirm students within the experimental group showed a significant increase in percentages of annotations being a response to fellow students after the intervention in week 3. We therefore cannot confirm our hypothesis that students who are scaffolded through collaboration scripts interact more often while performing their task in the SA tool, compared to students who are not scaffolded in this way.

Our second research questions was whether students, who are scaffolded through collaboration scripts, have a higher percentage of annotations showing higher order cognitive processing, on the levels of Bloom's revised taxonomy, while performing tasks on reading and annotating academic texts in a SA environment, compared to students who do not receive scaffolding through collaboration scripts while performing the same tasks? Our analysis showed that the percentage of annotations on the levels of higher order cognitive processing increased for the experimental group by 11.3% after the intervention. Though the percentage of the control group did decrease, this decrease (3.9%) was less severe as in our quantitative analysis and their scores stayed more consistent. The Mann-Whitney U test indicated that statistically significant differences, in favor of the experimental group, could be found for week 3 (after the intervention) when comparing the experimental and control groups. We also saw that, after the intervention, the experimental group scored higher percentages on the levels of 'Evaluating' (part of the higher cognitive processing levels) and lower on 'Remembering' (part of the lower cognitive processing levels) where the control group did not show a similar change. Finally, the Wilcoxon signed-rank tests showed a significant difference between weeks 2 and 3 for the experimental group, being a medium effect. Based on this analysis we can confirm our second hypothesis that students, who were scaffolded through collaboration scripts, more often showed levels of higher order cognitive processing, from Bloom's revised taxonomy, in their annotations than students who were not scaffolded in this way.

Our third research question was whether the effects of scaffolding through collaboration scripts on the difference of both the quantity and quality of interactions and annotations between students who received the scaffolding, compared to students who did not receive scaffolding while performing the same task, remain over time when the scaffolding of the first group is slowly faded out throughout the course? Looking at the percentages of annotations being interactions with fellow students, the Mann-Whitney U tests did show statistically significant differences in scores after the intervention and after the fading of the intervention, when comparing the experimental and control groups in weeks 3 and 6. However, looking at the results of the Friedman's ANOVA and Wilcoxon signed-rank tests, we could not confirm students within the experimental group showed a statistically significant increase in percentages of annotations, being a response to fellow students, throughout the three measurements over time. For the percentages of annotations on the levels of higher order cognitive processing we first of all saw that the increase after the intervention in week 3 for the experimental group was significant. However this percentage for the experimental group decreased again in week 6 to the same level of the control group. Though we did see the experimental group still scored higher on the level of 'Evaluating' and lower on the level of 'Knowledge' in week 6, the Mann-Whitney U test did not show a statistically significant difference between both groups for week 6. The Friedman's ANOVA and Wilcoxon test, for the experimental group, showed that the percentage of annotations on the levels of higher order cognitive processing of students did not stay significantly higher over time after the intervention. This means we cannot confirm our third hypothesis that, when the scaffolding of students through the use of collaboration scripts is faded out during the run time of a course, students who were supported through these scaffolds still show a higher number of interactions and a higher quality of interactions measured through Bloom's revised taxonomy levels, compared to students who were not scaffolded in this way.

Finally, though not part of our research questions, we were unable to validate the LICW2015-tool combined with verbs from Bloom's (revised) taxonomy as analysis instrument to score students annotations on different Bloom levels. Also, we wanted to see if we could find correlations between the scores of students on the percentage of interactions and their scores on the percentage of higher order cognitive processing from Bloom's revised taxonomy. We were however unable to find any correlations between these scores for both the experimental and control group.

6.3 Limitations, significance and future research

Only a small amount of research on SA tools has focused on supporting collaboration in learning tasks during assignments in these tools. Previous research on CSCL has shown that merely placing students in an online environment set up for collaborative learning, does not automatically mean they actually engage in collaboration or discussion. This study applied research from other areas of CSCL to the use of SA tools in higher education to investigate whether scaffolding students on a collaboration increases their engagement in discussions. From our findings we want to discuss a few limitations and recommendations for future research. First, because we could not validate the LIWC2015 tool in combination with the verbs/words list of the Bloom's revised taxonomy as instrument for scoring, we suggest that further research is required towards more complex and contextualized analysis. During our analysis we noticed that a consensus among the human raters, on how to assign student annotations on the levels of Bloom's revised taxonomy, needed to be reached to obtain medium to high correlations between the scores of the human raters. Furthermore, for our study, it also meant that the scoring had to be done by the researcher alone. Though the researcher did score the annotations without knowing which annotation belonged to what student, the fact that the researcher was scoring alone may have created bias. For future research we suggest scoring all annotations for analysis with a (larger) group of researchers while continuously checking for the inter-rater reliability scores.

In our discussion we also mentioned that a combination of instructions and feedback on collaboration during the assignments could encourage students to engage in collaborative behavior over a longer period in time. For future research we suggest studying and comparing the effects of different combinations of instructions (collaboration scripts) and feedback/feedforward during tasks in SA tools. Finally, as mentioned in our discussion, one element this study did not examine is to what extend the task in the SA tool itself and accompanying instructions influenced the willingness or motivation for collaboration amongst students. Given the limitations of this thesis we did not take this into account, however for future research into promoting collaboration during tasks in SA-tools we suggest researching the effects of differentiations in complexity of academic texts or the instructions of the assignments themselves. The mere selection of articles/reading material and the accompanying complexity and/ or relevance to the students as well as differentiation in the instructions may influence students' willingness to interact and discuss.

8. References

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston, MA: Allyn & Bacon (Pearson Education Group).

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York: David McKay.

Creswell, J.W. (2014*). Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research*. Essex, England: Pearson.

Dillenbourg, P., & Fischer, F. (2007). Basics of Computer-Supported Collaborative Learning. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 21*, 111-130.

Duffy, Th.M., & Cunningham, D.J. (1996). Constructivism: Implications of the design and delivery of instruction. In *D. Jonassen (Ed.), Handbook of Research for Educational Communications and Technology* (pp.170-195). London: Prentice Hall.

Ertmer, P., Sadaf, A., & Ertmer, D. (2011). Student-content interactions in online courses: the role of question prompts in facilitating higher-level engagement with course content. *Journal of Computing in Higher Education, 23*(2–3*),* 157–186.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock "N" Roll* (4th Edition). Los Angeles, London, New Delhi: Sage.

Gao, F. (2013). A case study of using a social annotation tool to support collaboratively learning. *The Internet And Higher Education*, *17*(3), 76-83. https://doi.org/10.1016/j.iheduc.2012.11.002

Ghadirian, H., Salehi, K., & Ayub, A. F. M. (2018). Social annotation tools in higher education: a preliminary systematic review. *International Journal of Learning Technology*, 13(2), 130–162. https://doi.org/10.1504/IJLT.2018.092096 .

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. New York & London: Routledge.

Kirschner, F., Paas, F., & Kirschner, P.A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology, 25,* 615–624. doi:10.1002/acp.1730.

Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hamalainen, R., Hakkinen, P., & Fischer, F. (2007). Specifying Computer-Supported Collaboration Scripts. *International Journal Of Computer-Supported Collaborative Learning*, *2*(2), 211-224. Retrieved October 24th 2018 from: https://link.springer.com/article/10.1007/s11412-007-9014-4

Kreijns, K. K., Kirschner, P. A., & Jochems, W. W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, *19*(3), 335–353. Retrieved November 5th 2018 from http://search.ebscohost.com.ezproxy.elib10.ub.unimaas.nl/login.aspx?direct=true&db=eoah&AN=4319510&site=ehost-live

Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal Of Computer-Supported Collaborative Learning (ijCSCL)*, *4*, 5-20. Retrieved November 5th 2018 from: https://hal.archives-ouvertes.fr/hal-00592750/document

Meyer, K. (2004). Evaluating online discussions: Four different frames of analysis. *Journal of Asynchronous Learning Network*, *8*, 101-114. http://dx.doi.org/10.24059/olj.v8i2.1830 .

Miller, K., Lukoff, B., King, G., & Mazur, E. (2018). Use of a Social Annotation Platform for Pre-Class Reading Assignments in a Flipped Introductory Physics Class. *Frontiers in Education*, *3* (8), 1-12. https://doi.org/10.3389/feduc.2018.00008

Mulcare, D. M., & Shwedel, A. (2017). Transforming Bloom's Taxonomy into Classroom Practice: A Practical Yet Comprehensive Approach to Promote Critical Reading and Student Participation. *Journal of Political Science Education*, *13*(2), 121–137. https://doi.org/10.1080/15512169.2016.1211017

Noroozi, O., Weinberger, A., Biemans, H. J. A., Mulder, M., & Chizari, M. (2012). Argumentation-Based Computer Supported Collaborative Learning (ABCSCL): A Synthesis of 15 Years of Research. *Educational Research Review, 7*(2), 79–106. https://doi.org/10.1016/j.edurev.2011.11.006

Novak, E., Razzouk, R., & Johnson, T.E. (2012). The educational use of social annotation tools in higher education: A literature review. *The Internet And Higher Education*, *15*(1), 39-49. https://doi.org/10.1016/j.iheduc.2011.09.002

Osborne, D. M., Byrne, J. H., Massey, D. L., & Johnston, A.N.B. (2018). Use of online asynchronous discussion boards to engage students, enhance critical thinking, and foster staff- student/ student- student collaboration: A mixed method study. *Nurse Education Today*, *70*, 40-46. https://doi.org/10.1016/j.nedt.2018.08.014

Prediger, S. & Pöhler, B. (2015). The interplay of micro- and macro-scaffolding: An empirical reconstruction for the case of an intervention on percentages. *ZDM Mathematics Education*, *47*(7), 1179-1194. http://dx.doi.org/10.1007/s11858-015-0723-2

Schellens, T., & Valcke, M. (2006). Fostering knowledge construction in university students through asynchronous discussion groups. *Computers in Education, 46*(4), 349–370, https://doi-org.ezproxy.elib10.ub.unimaas.nl/10.1016/j.compedu.2004.07.010

Sun, Y., & Gao, F. (2017). Comparing the use of a social annotation tool and a threaded discussion forum to support online discussions. *The Internet and Higher Education*, *32*(1), 72–79. doi:10.1016/j.iheduc.2016.10.001.

Rahman, S.A., & Manaf, N.F.A. (2017). A Critical Analysis of Bloom's Taxonomy in Teaching Creative and Critical Thinking Skills in Malaysia through English Literature. *English Language Teaching*, *10*(9), 245–256. doi:10.5539/elt.v10n9p245}

Valcke, M., De Wever, B., Zhu, C., & Deed, C. (2009). Supporting Active Cognitive Processing in Collaborative Groups: The Potential of Bloom's Taxonomy as a Labeling Tool. *Internet and Higher Education*, *12*(3), 165–172. http://dx.doi.org.ezproxy.elib10.ub.unimaas.nl/10.1016/j.iheduc.2009.08.003

Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-Cognitive Scaffolding with Computer Supported Collaboration Scripts: a Meta-Analysis. *Educational Psychology Review, 29*(3), 477-511. doi:10.1007/s10648-016-9361-7.

Wang, J., Wei, W., Ding, L., & Li, J. (2017). Method for analyzing the knowledge collaboration effect of R&D project teams based on Bloom's taxonomy. *Computers & Industrial Engineering, 103*(1), 158–167. https://doi.org/10.1016/j.cie.2016.11.010

Walter, S.D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in medicine, 17*(1), 101-110 https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1%3C101::AID-SIM727%3E3.0.CO;2-E

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning. *Instructional Science, 33*(1), 1-30. https://doi.org/10.1007/s11251-004-2322-4 .

Weinberger, A., & Fischer, F. (2006). A Framework to Analyze Argumentative Knowledge Construction in Computer-Supported Collaborative Learning. *Computers and Education, 46*(1), 71–95. http://dx.doi.org.ezproxy.elib10.ub.unimaas.nl/10.1016/j.compedu.2005.04.003

Winnips, K., & McLoughlin, C. (2001). Six WWW based learner supports you can build. In C. Montgomerie, & J. Viteli (Eds.). *Proceedings of the ED-MEDIA 2001 Conference* (pp. 2062-2068). Charlottesville, VA: AACE. Retrieved November 5th 2018 from https://eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED466231 .

Woods, K., & Bliss, K. (2016). Facilitating Successful Online Discussions. *Journal of Effective Teaching, 16*(2), 76–92. Retrieved October 25th 2018 from: https://eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=EJ1117812

9. Appendices

APPENDIX A

Collaboration script provided to students in the experimental group prior to assignment 3 on both a micro- and macrolevel.

Dear students,

After doing two assignments in Perusall you have built up some experience with this tool.
For the upcoming assignments we ask you to read the following instructions carefully and use them while working on your Perusall assignments. These instructions are meant to help you with these assignments:

In General: your Perusall assignments are not meant as just a task you 'check off'. The purpose is that by reading the articles and writing comments you gain a better understanding of the underlying concepts. What is important is that you can read and respond to questions and comments of a small group of your fellow students. To benefit from this, we ask you to not only focus on the text or on factual knowledge anyone can look up, but to engage in questions and discussions with your fellow students. Responding to your fellow students will not just help them, but also help you because explaining something to someone is a good way to (re)organize your own thinking and knowledge!

To help you with this we want to give you some suggestions you see below. We encourage you to use these suggestions.

Detailed suggestions:

- Out of your minimum of 9 comments, which are obligatory for the assignment, try to use **at least 3 as a response** to someone else's comment. If you are one of the first to do the assignment, this might mean you have to come back to check if you can make a response. Also, you could opt to do more than 9 comments, by later returning to add to other students comments. Keep in mind, there is no maximum limit! You can reply to as many comments as you like without it negatively influencing your grade on the assignment.

- When you write a comment on the text, instead of just commenting the text, try to start up a discussion. For this we want to provide you with alternative sentence starters you can use:

Instead of saying...	**Use:**
'In this segment the author connects to theory X'	'I would like to **argue** that the author connects to theory X or Y, **because…'**
'I am struggling to understand X or Y…..'	'Can anyone **help** me understand what the author means with X or Y? I struggle to understand this **because..**'
'This is a good/ bad argument the author is making.'	' I **feel** the author is making a good/bad argument here, **because**...'
'This is not how it was explained during the lecture!'	'Do you **agree** this seems to contradict with what was said during the lecture about X or Y…., **because…**?'

Important is that you try to give an argument for each comment or question you have, rather than just 'stating' something. This gives room for fellow students to respond to you. Also, be sure to check back if others have responded. If you follow up on someone else's comment, be sure to also use argumentation or try to use follow up questions to help the discussion along, instead of merely presenting a fact.

APPENDIX B

Collaboration script provided to students in the experimental group prior to assignment 4 on a macrolevel.

Dear students,

A couple of reminders of some of the key elements of these instructions for this weeks' assignment:

1. Remember that Perusall is about collaboratively discussing a text so you and your fellow students get a better grip and understanding of it.

2. Try to make a least 3 comments a response to comments of your peers.

3. Try to formulate your comments and responses in an argumentative way e.g. do not just 'state' something, but try to motivate it.

4. Don't start too late on your assignment, so you have time to check back and respond to any comments made towards your own remarks or to follow up on discussions.

The detailed instructions of last week are added in a word-doc if you want to view them again.

Good luck on the assignment!

The instructions with high intensity (instructions A) were added as attachment for students to read again if they choose to do so.

APPENDIX C

Verbs and words list based on Bloom's revised taxonomy, used to validate the LIWC2015-tool in this study. The highlighted words and verbs are present on more than one level.

Knowledge/ Remembering	**Comprehension/Understanding**	**Application/ Applying**	**Analysis/Analyzing**	**Evaluation/Evaluating**	**Synthesis/Creating**
		Apply	Analyze	Agree	Assemble
Define	Classify	Build	Assume	Appraise	Build
Find	Compare	Choose	Categorize	Assess	Choose
How	Contrast	Construct	Classify	Award	Combine
Label	Demonstrate	Develop	Compare	Compare	Compile
List	Explain	Interview	Conclusion	Conclude	Compose
Match	Illustrates	Make use of	Conclude	Criteria	Construct
Name	Infer	Model	Contrast	Criticize	Create
Omit	Interpret	Plan	Dissect	Decide	Delete
Recall	Relate	Select	Distinguish	Deduct	Design
Select	Rephrase	Solve	Divide	Defend	Develop
Spell	Summarize	Utilize	Examine	Determine	Elaborate
Tell	Translate	Act	Function	Disprove	Estimate
What	Asscociate	Back	Inference	Estimate	Formulate
When	Characterize	Back up	Inspect	Evaluate	Happen
Where	Clarify	Change	Motive	Explain	Imagine
Which	Convert	Complete	Relationship	Importance	Improve
Who	Discuss	Demonstrate	Simplify	Influence	Invent
Why	Distinguish	Discover	Survey	Interpret	Make up
Fact	Estimate	Dramatize	Take part in	Judge	Maximize
Concept	Express	Employ	Test for	Justify	Minimize
Answer	Extrapolate	Generalize	Theme	Mark	Modify
Identify	Generalize	Illustrate	Select	Measure	Original
Cite	Give	Interpret	Appraise	Opinion	Originate
Copy	Give	Manipulate	Break	Perceive	Plan
Describe	examples	Operate	Break down	Prove	Predict
Draw	Identify	Paint	Calculate	Rate	Propose
Duplicate	Indicate	Practice	Compare	Recommend	Solution
Indicate	Interpolate	Predict	Correlate	Rule on	Reorganize
Locate	Locate	Prepare	Debate	Select	Suppose

Memorize	Observe	Produce	Deduce	Support	Theory
Order	Paraphrase	Schedule	Detect	Value	Produce
Outline	Recognize	Show	Diagnose	Argue	Rewrite
Quote	Report	Simulate	Diagram	Arrange	Set up
Read	Represent	Sketch	Differentiate	Attach	Devise
Repeat	Restate	Use	Discriminate	Core	Facilitate
Recite	Review	Write	Experiment	Counsel	Generate
Recognize	Select	Administer	Figure	Critique	Hypothesize
Reproduce	Tell	Articulate	Group	Grade	Integrate
Review	Sense	Chart	Inventory	Manage	Make
State	Trace	Collect	Investigate	Mediate	Produce
Tabulate	Understand	Establish	Order	Probe	Rearrange
Underline	Visualize	Extend	Organize	Reconcile	Reconstruct
Enumarates	Subtract	Implement	Outline	Release	Revise
Record	Add	Include	Point out	Supervise	Role-play
Count	Approximate	Inform	Predict	Verify	Specify
Sequence	Factor	Operationalize	Prioritize	Weigh	Synthesize
Happened	Picture		Question	Reframe	Write
Meet	Correspond	Participate	Relate		Adapt
Study	Conceptualize	Project	Seperate		Anticipate
Tally		Compute	Subdivide		Collaborate
Score	Comprehend	Transfer	Focus		Incorporate
	Think	Acquire	Limit		Initiate
	Feel	Allocate	Corroborate		Reinforce
	Realize	Amend	Delegate		Structure
		Capture			Validate
		Conduct			Cultivate
		Convey			Overhaul
		Depreciate			
		Exercise			
		Expand			
		Graph			
		Transcribe			
		Put			

APPENDIX D

Instructions for the human raters for scoring students' annotations on the levels of Bloom's revised taxonomy.

Bloom Scoring instructions

1. Read the article students had to read for the week. This step is to ensure you understand the context of the material to some extent. This is also important because the scoring cannot be done in Perusall itself, so you cannot see what text the student selected. You will be able to see comments related to each other. You do not need to have a full understanding of everything in the article.
2. Open the data set. Read the comment carefully. Compare it to the definitions of Bloom's taxonomy. For help or reference you can also look at some of the example verbs and sentences for each level. However, keep in mind that some words and verbs may exist on different levels of Bloom. In the end make your own judgement on what level the text is on, mostly based on your interpretation of the definitions of the levels of Bloom's Taxonomy.
3. Give the **entire** comment **one** overall score (see schema).

The scoring schema is displayed below:

Context analysis schema for manual coding on the levels of Bloom's Taxonomy:

Category	**Definition**	**Example**	**Verbs**	**Score**
Knowledge/ remembering	The psychological process of remembering, recalling or recognizing facts, processes, structure. Exhibits previously learned material by recalling facts, terms, basic concepts and answers.	‘I recall the author of the book mentioned the same problem’ ‘This author was born in Scotland in 1972 as a child of wood smith’	Recall, State, Cite, Reproduce, Know	**1**
Comprehension/ understanding	Lowest level of understanding. When someone understands what is being represented. The understanding of the literal message of the material. Demonstrating understanding of facts and ideas by organizing, comparing, translating, interpreting, giving descriptions and stating main ideas.	‘If I understand correctly, the author is trying to make the point that…’ ‘From what I understand, the author is giving an explanation of...’…’	Illustrates, Characterize, Express, Comprehend, Understand	**2**

Application/ applying	If a question or problem is similar to questions or problems the student has faced before, the student demonstrates he or she can reapply previous knowledge or experiences to the current situation. I.e. once the meaning of the message is understood, it can be used to solve a new problem. Solving problems by applying acquired knowledge, facts, techniques and rules in a different way.	'The author uses a method here we could also use for…' 'The same principle can be applied to…'	Employ, Show, Apply, Use	**3**
Analysis/ analyzing	This demonstrates if a person can break down (or attempts to break down) a concept into its basic elements and identify relationships between those elements. Examining and breaking information into parts by identifying motives or causes; making inferences and finding evidence to support generalizations.	'From my own analysis, I see that there are four elements that define this problem that in turn all influence the main process, being….' 'The complexity of this problem can be broken down into different components. contributing to the problem, being…'	Dissect, Inspect, Correlate, Diagnose, Analyse	**4**

Evaluation/ evaluating	This is where judgements are made about ideas, work, materials and solutions and the value of those. Presenting and defending opinions by making judgments about information, validity of ideas or quality of work based on a set of criteria.	'Although I value your comment on this text, I want to argue you are neglecting some of the insights and concepts we talked about during last lecture, which could shed a different light on the topic...being…' 'I understand the position the author is taking here, based on X…..However I feel the author is neglecting some key elements from another theory which clearly states....'	Assess, Measure, Argue, Value	**5**
Synthesis/ creating	Creatively or divergently applying prior knowledge and skills to produce a new or original whole/principle/idea. Compiling information together in a different way by combining elements in a new pattern or proposing alternative solutions.	'Based on what we learned during last lecture and our discussion here, I want to propose a different solution to the discussion and problem,...'	Create, Improve, Propose, Synthesize	**6**