7th International Conference on Higher Education Advances (HEAd'21)
Universitat Politècnica de València, València, 2021
DOI: http://dx.doi.org/10.4995/HEAd21.2021.12862

The effects of collaboration scripts on the number and quality of student interactions in a social annotation tool

Vincent de Boer[1], Howard Spoelstra[2]

[1]Department of Educational Support and Innovation, University of Groningen, The Netherlands, [2]Faculty of Educational Sciences, Open University of the Netherlands, The Netherlands.

Abstract

Social Annotation (SA) tools can be used to facilitate active and collaborative learning when students have to study academic texts. However, making these tools available does not ensure students participate in argumentative discussions. Scaffolding students by means of collaborations scripts geared towards collaboration and discussion encourages students to engage in meaningful, high-quality interactions. We conducted an experiment with students (n=59) in a course running at a Dutch university, using the SA tool Perusall. A control group received normal instructions, while an experimental group received scaffolding through collaboration scripts. The results showed a significant increase in the number of responses to fellow students for the experimental group compared to the control group. The quality of the annotations, measured on levels of Bloom's taxonomy, increased significantly for the experimental group compared to both its baseline measurement and the control group. However, when scaffolding was faded out over subsequent assignments these differences became non-significant. The experimental groups' increased quality of annotations did not remain over time, suggesting that internalization of the scripts was not achieved.

Keywords: *Social Annotation; scaffolding; Bloom; collaboration scripts.*

1. Introduction

When preparing for lectures by reading academic texts, students' understanding of literature benefits from discussing it together (Miller *et al.*, 2018). Supporting such discussions can be achieved by using Social Annotation (SA) tools, allowing students to read academic texts online while sharing comments and questions (annotations), thus providing students with subject-oriented interactions (Sun & Gao, 2017). However, the fact that students have available such an environment does not mean they always participate in discussions (Kreijns *et al.*, 2003). When interaction is not stimulated, students may focus only on their own argumentation. Research on the use of discussion boards found that engagement in discussions benefits from scaffolding (Vogel *et al.*, 2017). It suggests that promoting interactions encourages students to engage in more meaningful discussions (Kreijns *et al.*, 2003). This scaffolding can be achieved by means of collaborations scripts (Vogel *et al.*, 2017), based on the theory that collaboration skills are internalized scripts, guiding students in the process of collaborative learning. The internalisation process is facilitated by using collaboration scripts (providing instructions and examples) that encourage learners to engage in argumentation and discussion (Vogel *et al.*, 2017). Script internalization can be expected, so scaffolding can be faded out which results in students self-directing their collaborative learning behavior without support. Research on collaboration scripts defines both high intensity micro-script scaffolding (containing suggested questions or sentence starters) and low intensity macro-script scaffolding (supporting meta-learning) (Kobbe *et al.*, 2007). In the next paragraph, we focus on how one can assess whether more meaningful interactions have actually occurred using Bloom's taxonomy (Bloom, 1956).

1.1 Bloom's taxonomy as instrument for measuring meaningful interactions in SA tools.

Bloom's taxonomy is a well-known instrument to assess whether students have actively processed information and applied cognitive skills such as comparing ideas and evaluating arguments. It defines six levels of cognitive processing, which can be bundled into lower (knowledge, comprehension and application) and higher (analysis, synthesis and evaluation) levels of deep learning (Bloom, 1956). These levels were reorganized by Anderson and Krathwohl (2001) into one of the lower levels (remembering, understanding, applying) or one of the higher levels (analyzing, evaluating and creating). In this study this taxonomy will be used to categorise the quality of student annotations into one of these levels and to assign them to either the lower or the higher level. For example, if a student wrote: 'If I understand correctly, the author is trying to make the point that…' we would score this on the level of Understanding (lower level) or if a student wrote 'I understand the position the author is taking here, however I want to argue that the author is neglecting key elements from theory X which clearly state....' we would score this on the level of Evaluating (higher level). We define three research questions:

RQ 1: Will students, who are scaffolded through collaboration scripts, engage in interactions more often while performing tasks in a SA environment, compared to students who do not receive scaffolding through collaboration scripts?

RQ 2: Will students, who are scaffolded through collaboration scripts, have higher percentages of annotations on levels of higher order cognitive processing of Bloom's revised taxonomy while performing tasks in a SA environment, compared to students who do not receive scaffolding through collaboration scripts?

RQ 3: Will effects of scaffolding through collaboration scripts remain over time when the scaffolding for the experimental group is slowly faded out during the course?

Based on the scaffolding-through-scripts theory and internalization theory, we formulate two hypotheses:

H 1: RQ 1 and RQ 2 will be confirmed. Students receiving scaffolding through collaboration scripts will show higher levels of interactions and a higher quality of annotations, measured with Bloom's revised taxonomy, compared to students who do not receive this scaffolding.

H 2: We expect that RQ 3 will be confirmed and that the higher levels of interactions and quality of annotations persist when scaffolding is faded out.

2. Method

An experiment was set up to analyse annotations students created in the SA tool Perusall during a Media Studies course of a Dutch university that ran in 2019. Over a period of 6 weeks students worked on one assignment per week. The targeted participants were all second-year students (n=102). Each assignment contained the instructions to read the prescribed literature and to create at least 9 annotations in the SA tool. The online learning environment randomly distributed the students over two conditions: control and experimental groups. Before assignment 1 both groups received the default Perusall instructions, explaining technicalities of the assignments and function towards classroom preparation. Before the second assignment the members of the experimental group were provided with collaboration scripts on both a micro- and a macro-level. These were communicated through both the electronic learning environment and by e-mail. After the third assignment scaffolding was faded out in two steps: partially before the fourth assignment and fully before the fifth assignment. The data to be analysed was collected from assignments 1, 2 and 5 and compared for differences on within-group and between-group levels.

2.1. Materials

The collaboration scripts we provided were derived from the previous research by Vogel *et al*. (2017), Noroozi *et al*. (2012) and Kobbe *et al*. (2007), consisting of macro-scripts explaining to students the importance of argumentation and interaction, and micro-scripts offering sentence starters to start discussions in annotations, such as: *'Instead of writing 'In this segment the author connects to theory X' we ask you to write 'I would like to argue that the author connects to theory X, because…''*.

2.2. Data collection and analysis

We collected data from three assignments: assignment 1 (for baseline measurement), assignment 2 (after the intervention) and assignment 5 (after scaffolding faded out fully). As students could skip one of six assignments, the number of students (n=59) providing data for all three measurement moments was lower than the number of students overall (n=102). The students for which we had full data sets showed an equal division over the experimental group (n=29) and control group (n=30). Each separate annotation received its own ID in the dataset. If an annotation consisted of a respons to another student, that annotation received an additional label, matching the ID of the original annotation (so-called Parent-ID). The interactions between students were labeled 0 when they were first annotations on some part of the text and 1 if annotations were a response to another student. We then calculated a response/first annotation ratio (response score) for each student per assignment on a range from 0 to1, for instance 0.2091 (20,91%). Second, we assigned the annotations to the various levels of Bloom's revised taxonomy. To validate these assignments to the levels, a sample of annotations (n=84) was assessed by three human raters (including the main researcher), assigning each annotation to a level of the taxonomy. Analysis showed moderate to strong inter-rater reliability. The complete data-set was scored by the main researcher, who assigned annotations considered to be on the lower-order levels a label of 0 and those on the higher-order levels a label of 1. We then calculated percentages of annotations per student for each assignment counting as higher-order cognitive processing on a range from 0 to1 (e.g. 0.350 or 35 % of annotations scoring on higher-order levels). From these we computed mean scores for each group from all annotations per assignment for interaction and Bloom-levels.

3. Results

3.1. Interaction between students

While comparing the mean scores (see Table 1) for assignments 1 and 2, we noticed that the experimental group's percentage of annotations as a response to fellow students increased. We also saw that, although decreasing, this mean was still higher for assignment 5 when compared to assignment 1. At the same time a decrease for the control group comparing their means of assignments 1 and 2 can be observed. Because our data showed a non-normal distribution, we combined Mann-Whitney U tests and a Friedman's ANOVA-test with a split file, followed by (post-hoc) Wilcoxon signed-rank tests to examine differences between assignments and groups, with a Bonferroni-correction of α (0.05/7)= .0071. First we checked for differences between groups per assignment. For assignment 1 the Mann-Whitney U test showed the median response score of the control group did not differ significantly from the median response score of the experimental group. This confirmed there were no differences between groups prior to the intervention.

Table 1. Means, Medians, Standard Deviations and Standard Errors of response scores.

Assignment	Group	Annotations	M	Mdn	SD	SE
1	Control (N=30)	N=245	.275	.222	.270	.049
	Experimental (N=29)	N=275	.335	.316	.258	.048
2	Control (N=30)	N=270	.180	.095	.233	.042
	Experimental (N=29)	N=266	.447	.444	.210	.039
5	Control (N=30)	N=277	.240	.222	.257	.047
	Experimental (N=29)	N=262	.391	.363	.222	.041

The experimental group's median response score of assignment 2 was significantly higher than that of the control group, U= 721, z= 4.375, p< .001, r=.57. For assignment 5 the experimental group's median response score was also significantly higher than that of the control group, U= 620.5, z= 2.832, p= .005, r=.37. The differences between groups, however, can be attributed to both the increased scores of the experimental group and the decreased scores of the control group. Unfortunately we could not verify the reason for the lower scores of the control group. The Friedman's ANOVA showed the percentages of annotations as response to fellow students for each group did not significantly change over time. The follow-up non-parametric Wilcoxon signed-rank tests showed no significant differences for each group between all assignments.

3.1. Scores on levels of Bloom's revised taxonomy

The scores from both groups on levels of Bloom's revised taxonomy (see Table 2) showed that both scored high on 'Remembering' for assignment 1 while scores on levels of higher order cognitive processing were relatively low. While the control group remained at these levels, the experimental group showed an increase of scores on higher levels for assignment 2, especially on 'Evaluating'. We noticed that both groups scored consistently high on 'Understanding'. This was not surprising as the texts were relatively new to the students. Next, we analyzed the percentages of annotations scoring on higher-order levels of Bloom's taxonomy. Here too, combined non-parametric tests (α = .0071) were required.

Table 2. Table of Bloom-scores for the experimental and control group.

	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
Experimental group						
Assignment 1 (n=275)	102	76	42	22	33	-
Assignment 2 (n=266)	67	75	32	21	70	1
Assignment 5 (n=262)	52	105	27	22	56	-
Control group						
Assignment 1 (n=245)	100	64	35	14	32	-
Assignment 2 (n=270)	135	68	26	13	27	1
Assignment 5 (n=277)	99	98	38	11	31	-

The results show that the scores for the experimental group increased from assignments 1 to 2 and, while somewhat lower, were still higher for assignment 5 (see Table 3). The scores of the control group showed a slight decrease when comparing assignments 1 and 2. The medians of percentages of annotations scored on the levels of higher order cognitive processing in assignments 1 and 5 did not differ significantly between both groups.

Table 3. Means, Medians, Standard Deviations and Standard Errors of levels of higher-order cognitive processing from Bloom's revised taxonomy.

Assignment	Group	Annotations	M	Mdn	SD	SE
1	Control (N=30)	N=245	.195	.163	.156	.028
	Experimental (N=29)	N=275	.201	.167	.200	.037
2	Control (N=30)	N=270	.156	.118	.122	.022
	Experimental (N=29)	N=266	.314	.333	.183	.034
5	Control (N=30)	N=277	.215	.222	.180	.033
	Experimental (N=29)	N=262	.237	.222	.175	.033

With assignment 1 being the baseline measurement, this confirmed there were no significant differences between the groups prior to the intervention. The median score of assignment 2 of the experimental group (Mdn= .333) was significantly higher than that of the control group, *U*= 671, *z*= 3.593, *p*< .001, *r*=.47. A Friedman's ANOVA test showed that the scores 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. The follow-up

Wilcoxon signed-rank tests for the experimental group showed a significant difference between assignments 1 and 2, *T* = 307.5 , *p*= .004, *r*= .37. We saw no significant differences for the experimental group between assignments 1 and 5 and between assignments 2 and 5 indicating the effect did not remain over time.

4. Conclusions, discussion and future research

Setting out to research the effects of collaborative scripting, our first hypothesis, suggesting that after scaffolding through collaboration scripts the experimental group will show higher levels of interaction and quality of annotations, could be partially confirmed. The results showed the experimental group did score significantly better on percentages of interactions compared to the control group on assignments 2 and 5. Our study also found a significant change in percentages of annotations categorisable as belonging to higher order cognitive processing in assignment 2 for the experimental group. For our second hypothesis we could not fully confirm that the effects of the scaffolding remained when scaffolding was faded out. Although there was still a difference between groups on levels of interaction on assignment 5, the increase in interactions was not statistically significant on a within-group level for the experimental group across the three assignments. Also, the control group's Bloom-score on assignment 5 was higher then their baseline score. This may be due to both groups showing natural progression throughout the course, not attributable to our intervention. Furthermore, the higher levels of annotation quality for the experimental group, measured with Bloom's revised taxonomy, did not remain when the scaffolding was faded out. A reason for this might be that collaborating requires an investment in time and effort and that the tasks or selection of texts were not complex enough to create a need to collaborate (Kirschner, Paas, & Kirschner, 2011) or the interdependence needed for collaborative learning (Johnson & Johnson, 1999). If the need to collaborate was low, this might also explain why these scripts were not sufficiently internalized to create a lasting effect. We measured the direct effects of macro- and micro-scripts while teacher feedback was only provided on the content-related effects based on the micro-scripts, and not on the collaboration-related performance based on the macro-scripts. Collaboration-related feedback might have stimulated the students in the experimental group to stay engaged in collaboration over a longer period of time and to internalize the corresponding behavior. This leads to two suggestions for future research: 1. To research the effects on the internalization of macro-scripts when tasks are designed with higher complexity levels, and 2. To research the effects of a combination of macro-scripts and teacher feedback on collaboration performance to enhance the internalization of the macro-scripts.

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. (Ed.) (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive Domain. New York: David McKay.

Johnson, D.W., & Johnson, R.T. (1999). Making cooperative learning work. Theory Into Practice, 38 (2), 67-73. https://doi.org/10.1080/00405849909543834.

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. https://doi.org/10.1007/s11412-007-9014-4.

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. https://doi.org/10.1002/acp.1730.

Kreijns, K.K., Kirschner, P. A., & Jochems, W.W. (2003). Identifying the pitfalls forsocial interaction in computer-supported collaborative learning environments: areview of the research. *Computers in Human Behavior*, 19(3), 335–353. https://doi.org/10.1016/S0747-5632(02)00057-2

Miller, K., Lukoff, B., King, G., & Mazur, E. (2018). Use of a Social Annotation Platformfor 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.

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

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

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. https://doi.org/10.1007/s10648-016-9361-7.