Authors

Mario Gielen: http://www.tecolab.ugent.be/pages/mario.html

Bram De Wever: http://www.tecolab.ugent.be/pages/bram.html

Structuring the Peer Assessment Process: A Multilevel Approach for the Impact on Product Improvement and Peer Feedback Quality

Abstract

In order to optimize students' peer feedback processes, this study investigates how an instructional intervention in the peer assessment process (PA) can have a beneficial effect on students' performance in a wiki environment in first-year higher education. The main aim was to study the effect of integrating a peer feedback template with a varying structuring degree. The present study involved three conditions: a no structure, a basic structure, and an elaborate structure condition. Due to a clear hierarchical structure, in which over time (level 1), 168 students (level 2) are nested within 37 groups (level 3), multilevel analysis was performed to examine the effect of time, student and group level influences on students' peer feedback quality and product scores. Results revealed that both peer feedback quality and product scores increase significantly for all conditions over time, after multiple practice occasions. In addition, after several practice occasions, significant differences were found between the conditions in both peer feedback (elaborate higher than no structure) and product scores (elaborate and basic higher than no structure). Building on this, limitations, directions for future research, and practical implications are presented.

Keywords

Peer assessment, peer feedback quality, structuring, higher education

Introduction

Peer feedback as educational practice

Several authors (eg. Black & William, 1998) have emphasized on the power of assessment for the learning process, rather than assessment of learning. Peer assessment (PA) is one specific method that can be employed for formative assessment and thus reach this aim while involving learners in all phases of the assessment process (Dysthe, 2004). Previously, peer assessment is defined as "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal- status learners" (Topping, 2009, p. 20). Therefore, PA is often suggested as a good approach for increasing students' engagement with their own learning (eg. Nicol & MacFarlane-Dick, 2006). To this day however, research on PA in higher education remains "very variable in type and quality, scattered and fragmentary in nature" (Topping, 1998, p. 267; see also Evans, 2013, who presents the same conclusion).

In the context of PA, peer feedback is often perceived as an educational activity for enhancing students' learning (eg. Falchikov, 1995), in which peers juggle with "information provided by an external agent regarding some aspect(s) of the learner's task performance, intended to modify the learner's cognition, motivation and/or behaviour" (Duijnhouwer, Prins, & Stokking, 2012, p. 171). Previous research pointed out that peer feedback enhances students' performance (eg. Falchikov, 2003). In view of formative assessment, it is rather logical that students should be given the opportunity to use this feedback, in order to improve their learning and achievement (Nicol & MacFarlane-Dick, 2004). However, many questions remain unanswered on how the formative assessment practices should be implemented into educational practice to boost students' learning in higher education (Sadler, 2010). More particular, research lacks a rigid approach on how PA practices should be tailored in function of students' learning

(Strijbos & Sluijsmans, 2010). In this respect, this study is particularly focusing on how instructional interventions can customize the PA process in order to enhance students' learning.

The essence of peer feedback quality

A growing body of research emphasizes that feedback has a powerful impact on both learning and performance (Nelson & Schunn, 2008). Interestingly, the average effects of feedback are one of the highest in education, but also one of the most unpredictable in their influences (Hattie & Gan, 2011). A large body of research claims that the effectiveness of a feedback message largely depends on the content, form and function of the feedback (eg. Narciss, 2008). Especially, feedback content appears to be crucial for the impact of feedback on learning and performance (Cho & MacArthur, 2010). In literature, there is no fixed answer on what exactly determines peer feedback quality. Following the feedback framework of Hattie and Timperley (2007), high quality peer feedback should provide answers on three major feedback questions: 'Where am I going?', 'How am I going?', and 'Where to next?'.

While other studies propose to examine the quality of peer feedback messages through content analysis (eg. Strijbos, Van Goozen, & Prins, 2012; Gielen & De Wever, 2013), or through the calculation of two indices, namely validity and reliability (Hafner & Hafner, 2003), prior studies have applied a scoring rubric to measure the quality of the feedback messages (Prins, Sluijsmans, & Kirschner, 2006). A scoring rubric is particularly valuable because it presents the assessment criteria in a structured format (Panadero, Romero, & Strijbos, 2013) and it gives an indication about expected performance by listing the relevant assessment criteria and by defining the quality levels of each criterion (Andrade & Valtcheva, 2009). Prins et al. (2006) developed the Feedback Quality Index, in which a number of quality criteria are discussed. First

of all, they emphasized on the importance of the assessment criteria, in which assessor and assessee are guided towards high quality performance (see also Sluijsmans, 2002). This idea is also supported by research, which claims that it is essential that the assessor is capable of identifying and understanding the assessment criteria in order to provide a reliable and valid assessment (Panadero & Jonsson, 2013). Next, students must have the knowledge and skills to sufficiently illustrate the nature of their feedback. In the FQI (Prins et al., 2006), the nature of feedback refers to specific peer feedback content such as remarks, posed questions and external examples. Logically, previous research suggested that some types of feedback are more effective than others (Nelson & Schunn, 2008). Previous research revealed that more specific and elaborated feedback leads to improved performance and outcomes (Strijbos, Narciss, & Dünnebier, 2010). Finally, students need to be capable to transform their peer feedback in a message. According to the FQI (Prins et al., 2006), students should write their feedback in the first person throughout the whole report, in a logical and clear structure, in which short descriptions are preferable. (Prins, et al., 2006).

In an attempt to safeguard high quality peer feedback, recent research summarizes that students are involved in high-level cognitive processing during this peer feedback process (King, 2002), in which they require skills comprising "the ability to engage with and take ownership of evaluation criteria, to make informed judgments about the quality of the work of others, to formulate and articulate these judgments in written form and, fundamentally, the ability to evaluate and improve one's own work based on these processes" (Nicol, Thomson, & Breslin, 2014, p. 120). Therefore, this study is particularly interested in how we can optimize the peer

feedback process with the underlying purpose to increase the feedback quality and additionally, the product score.

Structuring the peer feedback process to optimize feedback quality

As mentioned above, PA can be seen an example of a more complex learning task that requires high-level cognitive processing, however, such high-level PA processes hardly happen spontaneously (Kollar & Fischer, 2010). Previous studies pointed out the need for structure and support to ensure effective feedback (eg. Poverjuc, Brook, & Wray, 2012). Recently, research questioned what type of support is essential for the assessor and assessee to promote high quality feedback (Hovardas, Tsivitanidou, & Zacharia, 2014). Previous research of Van Merriënboer, Kirschner, and Kester, (2003) suggested amongst others to model the use of cognitive strategies by presenting checklists and process worksheets, or by asking leading questions, in order to support students in complex learning. This type of support may be beneficial to support the role of the assessor in providing feedback as well.

Other studies showed that structure is beneficial for the peer feedback process by, for example further specifying a peer feedback template to enhance the peer feedback quality (eg. Gielen & De Wever, 2012). It is within this frame that the main aim of the present study can be situated: “How can we increase the peer feedback quality by structuring the PA process?” Based on the scripted cooperation approach (O’Donnell, 1999), collaboration scripts are recommended in the literature to boost successful collaborative learning activities (Fischer, Kollar, Stegmann, & Wecker, 2013). As a script specifies, plans, and assigns roles and activities for collaborative learning activities (eg. Fischer, et al., 2013), a script can be seen as an instructional collaboration scenario (O’Donnell & Dansereau, 1992), which concentrates on socio-cognitive structuring

(Kollar, Fischer, & Hesse, 2006). Since numerous contextual factors play a role, determining the accurate level of structuring appears to be the actual challenge (Dillenbourg, Järvelä, & Fischer, 2009). A previous study showed that structuring the peer feedback process by providing a peer feedback template consisting out two guiding questions, in which the first one focuses on providing feedback and the other one focuses on providing feed forward, leads to significant higher peer feedback and product scores (Gielen & De Wever, 2012). For this reason, the elaborate structure in the peer feedback template goes a step further, as it is organized according to the feedback principles of Hattie and Timperley (2007), for each criterion separately. As no previous studies investigated the impact of a similar peer feedback template with a varying structuring degree based on these feedback principles, this study attempts to provide an answer on how detailed the script should be and what level of structuring is the most appropriate (c.f. 'script granularity' concept of Kobbe, et al., 2007), taking into account under-scripting (Kirschner, Sweller, & Clark, 2006) or over-scripting effects (Dillenbourg, 2002), in which a script can be too flexible or too rigid that it eventually undermines students' learning.

In order to become skilled peer assessors and assessees, who provide and receive high quality peer feedback, research stresses that students require practice and training (Sluijsmans, 2002; Birenbaum, 1996). As training is often suggested in the literature, it is important that students have the opportunity to practice similar performance at multiple occasions. For this reason and building on previous studies (eg. Gielen & De Wever, 2012), this study foresees three performance cycles. This is in line with research, which claims that students need to have the opportunity to replicate similar performance or to close the feedback loop, in order to grasp the effectiveness of the peer feedback (Boud, 2000). While performing an academic task in a wiki

environment, which is praised for supporting online collaboration and assessment activities (De Wever, Van Keer, Schellens, & Valcke, 2011), this study incorporated three different feedback forms with a varying structuring degree as instructional intervention, to examine the effect on the feedback quality and product score.

Rationale for this study and expectations

It is within this frame that this study is particularly interested in to what degree the assessors' peer feedback process should be structured, in order to increase the peer feedback quality and product scores. With respect to this question, we expect peer feedback quality scores and the product scores will increase over time, as mentioned above, when learners have the opportunity to perform similar tasks at multiple measurement occasions. As students in their bachelor program habitually lack practice and experience in the peer feedback process, they may require a higher amount of structure and support, in order to become skilled peer assessors who provide high-quality peer feedback. Therefore, we expect that a higher structuring degree will lead to higher peer feedback quality scores. Consequently, we assume that students, who receive a higher degree of structuring in their peer feedback process, will have higher product scores and a higher increase from draft to final version, compared to less-structured conditions.

Methodology

Participants and procedure

The participants in the present study were first-year bachelor students Educational Sciences (N = 168), enrolled in the course Instructional Sciences that runs during the first semester of the academic year. Participants were randomly assigned to groups (n = 37) of

maximum 5 students to collaborate in a wiki environment. Students had to write three times a draft and final version of an abstract of a submitted, yet not published scientific article related to the topic (ie. they received the paper, but the abstract was left out). Before writing the final version, they received peer feedback on their draft version, formulated on a provided peer feedback template. Each student was assigned to provide three times peer feedback (one time for every one of the three draft versions written) to one fixed specific group member with the goal to increase the quality of the final abstract. Regarding the amount of time given for the assignment, students had a week time for each particular step, which adds up to a 3 week period for writing a draft version, providing and receiving peer feedback, and writing a final version of an abstract. As students were involved 3 times in this cycle, the total amount of given time for the assignment was 9 weeks. The wiki task, including the peer feedback, was part of their curriculum requirements. During the writing and assessment phase, students could access the wiki anywhere and anytime.

Research instruments

Scoring rubric for quality of peer feedback messages

First of all, the rubric to assess the peer feedback quality is based on the Feedback Quality Index (Prins et al., 2006), which is in turn based on several prior studies (eg. Sluijsmans, Brand-Gruwel, & Van Merriënboer, 2002). Following the scoring rubric that was developed to measure the quality of feedback reports of general practitioners in training (Prins et al., 2006), other previous studies (eg. Gielen & De Wever, 2012) and this particular study applied a scoring rubric, which maintains all three main categories (use of criteria, nature of the feedback, and writing style), and their involved sub categories with corresponding scoring percentages of the

scoring rubric of the FQI, but focused specifically on measuring the quality of peer feedback messages of first-year higher education students. First of all, Use of criteria was categorized by the number of: used criteria, remarks per criteria, remarks focused on particular aspects of criteria, explanations of remarks per criteria, explanations of remarks focused on particular aspects of criteria. Similar to the FQI, the use of criteria accounted for 50% of the score, in which both feedback content and explanations were assessed. Secondly, nature of feedback was categorized by the number of: positive and negative remarks, reflective questions, external examples and suggestions for improvement. These four items that identify the nature of feedback accounted in total for 35%. Finally, writing style was categorized by: structure, use of key words or descriptions, and use of first person (Prins et al., 2006). These three items defined the quality of writing and accounted for 15%. As shown in Table 1 in the appendix, this resulted in a scoring rubric of 9 items with a scoring range between 0 and 100 to measure the quality of peer feedback messages.

Scoring rubric for quality of the wiki product

For the product score, ie. the quality of the written abstract in the wiki, a rubric was developed in which the necessary components for a good abstract are included. In academic writing, literature refers to components such as intention of the study, problem statement, methodology, findings and conclusions, limitations, structure, language, etc. Therefore, this study developed a scoring rubric in which these components are categorized in four main categories (situating the study, content of the abstract, style, and general impression) and nine corresponding sub categories. First of all, situating the study was categorized by how well described are: the intention or focus of the study, the context of the problem statement, and

finally the continuity between the focus of the study and the context of the problem statement. These three items accounted for 30% of the total score. Secondly, the content of the abstract was categorized by the methodology with corresponding details on the setting, the results being all present and concisely formulated, and finally the presence of limitations and suggestions for future research. These three items accounted for 25% of the total score. Thirdly, the main category style was categorized by: structure of the abstract, language and writing style, and finally word count. These three items accounted for 25% of the total score. Finally, the main category 'General impression' was categorized by the impression of completed effort and corresponding need for revision. This main category accounted for 20% of the total score. Therefore, the developed scoring rubric to analyze the quality of the wiki product had a scoring range between 0 and 100, as shown in Table 2 in the appendix.

Conditions

The instructor provided a peer feedback form for each of the three conditions, presented as a template with a list of ten criteria (intention of research, problem statement, methodology, results, conclusion, limitations, structure, language, deadline, and general judgment). This list of criteria was submitted to the no structure condition, but students in this condition received no further instructions, while the two other conditions received additional instructions. The basic structure condition received additionally two guiding questions. First of all, students were directed to provide feedback on how well their peers performed in past performance, by answering the following question: "What was good about your peers' work?" Secondly, students were encouraged to provide feed forward, in which suggestions could be made in function of future performance, by answering the following question: "What would you change in your peers' work?'). By receiving a higher structuring degree, students in the elaborate structure

condition incorporated a peer feedback template, which was structured according the principles of feed up, feedback, and feed forward (Hattie & Timperley, 2007) and additionally in each of these three sections, the list of criteria was simply repeated. Since students were instructed to follow these three steps, they firstly started with formulating feed up for each of the ten criteria. An example of feed up for the criterion 'problem statement' could be that the assessee needs to explain more concise the problem in relation with the research intention. Secondly, students had to provide feedback on the ten criteria, in which they for example explain how well the assessee described the problem statement. Thirdly, students were instructed to formulate feed forward for each criterion, in which they could suggest for example to add more details to the problem statement, as a suggestion for future improvement.

Data analysis

Given the clear hierarchical structure of the data, namely three measurement occasions (i.e. the peer feedback moments, indicated by the variable 'time', level 1) are nested within each of the 168 students (level 2), who are in turn nested within 37 groups (level 3), multilevel modelling (MLwiN 2.29) was used to analyze the peer feedback quality and the product quality (ie. the quality of the versions of the abstract written in the wiki).

Initially, for both peer feedback score and product score a fully unconditional null model was tested to examine whether a multilevel approach was required compared to a single-level regression analysis. Next, the categorical predictor time was added to the null model, which resulted in a compound symmetry model, which is a random intercept model with no explanatory variables except for the measurement occasions (Snijders & Bosker, 1999). In this model, the two last measurement occasions (time 2 and 3) can be compared with the reference category (time 1). After this, the followed procedure is dissimilar for the peer feedback and product score.

Regarding the peer feedback score, the categorical predictor ‘condition’ is added in the next step. In a final phase, the interaction condition*time was added to the model. Regarding product score, first the categorical predictor ‘version’ was added to the model, as the product score has two versions, namely draft and final version. After this, the categorical predictor condition is added in a next phase. Finally, the interaction time*condition was added to the model. Further exploration of other interaction possibilities revealed no significant interaction effects when version was involved and was therefore excluded from the multilevel model. By using a stepwise multilevel approach, the additional value of each subset of variables to the model could be checked.

Hypotheses

With respect to the quality of the peer feedback, the following hypotheses are put forward:

(H1) Over time, students in all three conditions will have significantly higher peer feedback quality scores, more specifically (H1a) from time 1 to time 2, (H1b) from time 2 to time 3, and (H1c) from time 1 to time 3.

(H2) Students will have higher peer feedback quality scores, more specifically (H2a) the basic structure compared to the no structure condition, (H2b) the elaborate structure compared to the no structure condition, and (H2c) the elaborate structure compared to the basic structure condition.

With respect to the quality of the product, the following hypotheses are put forward:

(H3) Over time, students in all three conditions will have significantly higher product quality scores, more specifically (H3a) from time 1 to time 2, (H3b) from time 2 to time 3, and (H3c) from time 1 to time 3.

(H4) Students will have higher product quality scores, more specifically (H4a) the basic structure compared to the no structure condition, (H4b) the elaborate structure compared to the no structure condition, and (H4c) the elaborate structure compared to the basic structure condition.

(H5) The product quality scores improve significantly better from draft to final version for students, more specifically (H5a) for all conditions, no matter what structuring degree they receive (main effect), but even more in (H5b) the basic structure compared to the no structure condition (interaction effect), (H5c) the elaborate structure compared to the no structure condition (interaction effect), and (H5d) the elaborate structure compared to the basic structure condition (interaction effect).

Results

Peer feedback score

All models were created following the previously described stepwise procedure and are represented in Table 3 in the appendix. The random-intercept three-level null model (Model 0) predicts the overall peer feedback score across all feedback moments (time), students, and groups (the intercept; ie. 53.23 out of 100). The null model divides the variance of peer feedback scores into between groups, within groups between students, and within students between peer feedback moments. The results show that 19.32% of the total peer feedback variance is situated

at the group level (p=.002), the proportion of variance due to differences between students within groups was 11.23% (p=.014), and finally 69.45% of the total variance is situated at the time level (p<.001, see Table 3).

Next, the categorical predictor time was added to the null model, which resulted in Model 1. Adding this variable to the null model resulted in a better model fit ($\chi 2$=98.309, df=2, p<.001). The results presented in Model 1 reveal a significant effect of measurement occasion on peer feedback scores, indicating that the quality of the feedback was significantly higher the second and the third moment (compared to the first moment, both at p<.001). On average, the quality of the feedback was also significantly higher the third moment compared to the second (p=.032). Following, the categorical predictor condition was added to Model 1, which resulted in Model 2. The condition in which students did not receive any additional structure in their peer feedback template was taken as reference category. Adding this variable resulted in a better model fit ($\chi 2$=13.308, df=2, p=.001). In the last step, the interaction effects between time and condition were added. However, as this model did not result in a better fit than Model 2 ($\chi 2$=1.605 df=4, p=.808) and none of the interaction effects were significant, Model 2 was chosen as final model for further analysis.

In Model 2, the results indicate a significant main effect of measurement occasion and condition, with respect to the peer feedback scores, as shown in Figure 1. First of all, peer feedback scores increased significantly over time for all students in all groups, both significantly from time 1 to time 2 with an increase of 13.12 (p<.001), and from time 2 to time 3 with an increase of 3.51 (p=.033), causing a total increase from time 1 to time 3 of 16.625 (p<.001). These findings confirm respectively H1a, H1b and H1c. Secondly, results point out that students

who received an elaborate structure have an overall significantly higher peer feedback score, which is in more detail 11.79 higher compared to the no structure ($p=.001$), and 13.51 higher than the basic structure ($p<.001$) condition, confirming H2b and H2c. Between the no structure and basic structure no significant differences were found ($p=.190$), not supporting H2a.

Please insert Figure 1 here

Product score

In Table 4 in the appendix, the null model showed that 1.02% of the total peer feedback variance is situated at the group level ($p=.478$), the proportion of variance due to differences between students within groups was 5.58% ($p=.040$), and finally 93.40% of the total variance is situated at the time level ($p<.001$). After estimating the null model, the categorical variables time and version were added to the null model as measurement occasions (Model 1). The results presented in Model 1 reveal a significant main effect of measurement occasion on product scores over time. Adding these two variables resulted in a better model fit ($\chi2=838.695$, $df=3$, $p<.001$). After estimating Model 1, interaction effects between time and versions were checked for, but no significant effects were found, indicating that the increase in score between the draft version and the final version was about the same at each of the three moments. For Model 2, the categorical predictor condition was added, and the results revealed a significant main effect. Model 2 did not fit the data better than Model 1 ($\chi2=0.626$, $df=2$, $p=.731$) and no main effect of condition was found (see Model 2, Table 4). After estimating Model 2, interaction effects between condition and version were checked for but not found. However, in a next step, the interaction effects of

time and condition were added to Model 2, resulting in Model 3, and revealing significant interaction effects. Model 3 also fitted the data better than both Model 2 ($\chi 2$=20.884 df=4, p=.001) and Model 1 ($\chi 2$=21.510, df=6, p=.001), and was therefore chosen as final model for further analysis.

In Model 3, results indicate a significant main effect of the two categorical predictors time and version, with respect to the peer feedback scores. Firstly, results show that the product scores increased significantly over time for all students, confirming H3a, H3b and H3c. More specifically, the product scores improved significantly from time 1 to time 2 with an increase of 22.439 (p<.001), as well as from time 2 to time 3 with an increase of 7.18 (p<.001), causing a total increase from time 1 to time 3 of 29.61 (p<.001). Secondly, results point out that the product score increased significantly from draft to final version for all students, confirming H5a, with an average increase of 9.14 (p<.001). No interaction effects were found between condition and version and therefore H5b, H5c and H5d are not supported.

Regarding condition, multilevel analysis revealed no main effect, contradicting H4a, H4b and H4c. Though when taking into account the feedback moments, results showed an interaction-effect, suggesting that the product scores evolve significantly different over time for particular conditions, as shown in Figure 2. More specifically for time 1, results show that the product score of the basic condition was 0.70 lower compared to the no structure (p=.778). Students in the elaborate structure condition had a lower product score at the start, which is in more detail 4.73 lower compared to the no structure (p=.063), and 4.03 lower compared to the basic structure condition (p=.114) at time 1.

Regarding time 2, results reveal that students, who received an elaborate structure, have a slightly higher (but not significant) product score, which is in more detail 1.89 higher compared to the no structure (p=.457), and 1.42 higher compared to the basic structure (p=.573) condition. Also the product score of the basic structure was 0.47 higher compared to the no structure condition (p=.851) at time 2. Finally for time 3, results show that students who received no structure have an overall significant lower product score, which is in more detail 5.14 lower compared to the basic structure condition (p=.039), and 5.78 lower than the elaborate structure condition (p=.023). This only partly (i.e. only at time 3) confirms H4a and H4b. The product score of the basic structure was 0.64 lower compared to the elaborate structure condition (p=.802), so H4c is not supported.

Please insert Figure 2 here

In sum, Figure 2 represents the main findings clearly. Firstly, the progress from draft to final version is almost equal for all conditions at all moments (see the similar slopes in Figure 2). Over time, product scores improve overall, but point out no differences between the conditions. However, if we look closer at each moment, results show that at time 1 the elaborate structure has lower (but only nearly significant) product scores than the two other conditions, while at time 2 the elaborate structure already has slightly higher, but not significant higher product scores compared to the less structured conditions. Interestingly, at time 3, both the elaborate and basic structure condition have significantly higher product scores compared to students who receive no additional structure in their feedback process.

Discussion

This study examined how the degree of structuring the peer assessment process has an impact on the feedback and product quality, when students compose feedback with the help of a peer feedback form with a varying structuring degree. Finally, the practical implications and direction for future research are presented.

Peer feedback quality

Over time, the results revealed that the feedback scores increase significantly for all conditions, suggesting that students overall provide peer feedback messages of a better quality, when they gain experience through practice. Although a few studies claim that students can offer valuable feedback without actual training in assessment (eg. Cho & MacArthur, 2010), other research advocates that students benefit from practice and training in receiving and providing peer feedback (eg. Sluijsmans, 2002) and moreover, that students require practice to become skilled peer assessors (eg. Birenbaum, 1996; Van Steendam, Rijlaarsdam, Sercu, & Van den Bergh, 2010). Measured by the Feedback Quality Index (Prins et al., 2006), the results of the present study revealed that students in the elaborate structure condition have significantly higher feedback quality scores, compared to students who received merely some guiding questions or who received no additional structure at all. This is in line with previous experimental studies, which claim that structure is beneficial for the peer assessment process (eg. Gielen & De Wever, 2012) and, which underline the need for structure and support to ensure effective feedback (eg. Poverjuc, Brook, & Wray, 2012). More specifically, offering students a peer feedback form including a criteria-oriented list structured according the three feedback principles feed up,

feedback and feed forward (Hattie & Timperley, 2007) appears to be an effective approach to increase significantly the peer feedback quality. As finding the accurate level of scripting is the actual challenge (Kobbe, et al., 2007), we believe that further scripting the peer feedback process by providing an elaborate structure in a peer feedback template is a favorable approach, to enhance the peer feedback quality (eg. Fischer, Kollar, Stegmann, & Wecker, 2013).

Product quality

The results indicated an overall significant increase over time for all students, no matter what level of structure they receive in their peer feedback process. By engaging students actively in PA, previous research claims students' learning can be facilitated (eg. Li, Liu, & Steckelberg, 2010), as PA has several cognitive gains for both assessor and assessee, such as increased attention on the crucial elements, which determine high quality work (Topping, 1998). Following, results demonstrated an overall significant increase of product scores of 9% from draft to final version. This is in line with research, which underlines that feedback can have a large impact on performance (Nelson & Schunn, 2008), as it "might also reveal the next small steps needed to improve quality" (Topping, 1998, p. 255). This is supported by a review, which advocates that every variety of feedback, whatever its amount or specificity, can have a positive effect on students' product scores (Topping, 1998).

With respect to the provided level of structure in the peer feedback process, overall results revealed no significant differences between the conditions regarding product quality scores. However, when taking a closer look, interaction effects pointed out some significant differences between the conditions over time. In general, research advocates that the quality of students' performance increases over time, whenever they have the opportunity to practice

similar learning activities (eg. Sluijsmans, 2002). More specifically in this study at time 3, results indicated that students of both basic and elaborate structure conditions had significantly higher product quality scores after multiple practice occasions, compared to students who did not receive additional structure in the peer feedback process. These findings suggest that structure in the peer feedback process has the potential to boost product scores, while it is important that students use this feedback, in order to improve their performance (Nicol, & MacFarlane-Dick, 2004). This is supported by other research, which advocates that structure, in which the roles and activities of involved learners are further concretized, can be valuable for students' learning (Schellens & Valcke, 2006). It is important to notice that other research has shown that peer feedback does the not necessarily increases the quality of performance over time, especially not in a later phase of PA activity (Chen & Tsai, 2009). While the present study showed an effect in the later phase, earlier research (Tseng & Tsai, 2007) showed that suggestive feedback was especially valuable in the initial phase of PA, but that its importance declined in later phase. Future research is necessary to shed more light on when exactly students benefit the most from these activities.

To conclude, our results indicate that further scripting the peer feedback process can be beneficial for the quality of students' peer feedback and product performance, which is in line with similar previous studies (eg. Gielen & De Wever, 2012). Over time, all students improved significantly after multiple practice occasions in providing peer feedback and finishing their wiki task. It became clear that in the end students, who received an elaborate degree of structure to provide peer feedback, had significantly higher peer feedback quality scores compared to less structured conditions. Furthermore, students who received additional structure in their peer

feedback template in the end had significantly higher product quality scores after similar practice occasions, compared to students who did not receive any additional structure in the peer feedback process. Therefore, this study advocates that offering additional structure in PA, to further specify the role of the assessor during the peer feedback process, is a valuable approach to increase both the quality of peer feedback and performance.

Limitations, directions for future research and practical implications.

The present study took place in an authentic learning environment. While this is a large advantage in view of creating an ecologically valid setting, it has the disadvantage that not all contextual factors can be controlled. While we performed a manipulation check, i.e. we (1) checked whether students provided feedback, (2) whether they used the assigned template, and (3) whether they read and used the peer feedback (they had to indicate sentences in color that were changed based on the feedback), to ensure treatment confidence, there is still a possibility that other factors, such as maturation or studying course content throughout the semester, can have had an impact with respect to the increase in product scores from draft to final version. Given that the task of writing an abstract is a specific task and these competences are not studied in the curriculum during that time, we expect low influence of other contextual variables, however, we have to take into account that a significant increase of performance could be not necessarily the result of the received peer feedback, but as well the result because of maturation (Kluger & Denisi, 1996).

Future studies could close the feedback loop (Boud, 2000), in which the assessee could be structured to evaluate the received peer feedback after revision. Moreover, the present study incorporated the FQI to measure the quality of peer feedback messages, while in-depth content

analysis could be another approach to determine the actual peer feedback quality, as it provides a more detailed insight of the specific peer feedback content. Another direction for future research could be to examine the added value of a structured peer feedback form in various educational contexts and over a longer period of time. Also, it could be valuable to investigate when, and at which time exactly, peer feedback is most effective to increase the quality of performance.

A last suggestion for future research could be to examine how the role of the assessee could be structured as well, by for example a peer feedback request, whilst the majority of the experimental studies in the literature focus merely on the assessor (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010).

Instructors wishing to implement peer assessment should consider the following two recommendations. Firstly, this study recommends implementing a peer feedback template with a higher structuring degree, as instructional intervention to support students during the peer feedback process. Therefore, this study proposes that a peer feedback template should consist out of two essential features. On one hand, the template needs to provide a list of the pre-specified, or preferably mutual discussed criteria (Sluijsmans, 2002). On the other hand, this template could be inspired by feedback framework of Hattie and Timperley (2007), in which students are encouraged to provide feedback on past performance and feed forward in function of future performance, focused on particular criteria. Secondly, this study supports a large body of research that encourages instructors to foresee multiple practice occasions, in which students are involved in peer assessment and similar task performance.

References

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice, 48*, 12–19.

Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum, & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-29). Boston, MA: Kluwer.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*, 7-74.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society Studies in Continuing Education, 22, 151–167.

Chen, Y. C., & Tsai, C. C. (2009). An educational research course facilitated by online peer assessment. *Innovations in Education and Teaching International, 46*, 105−117.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*, 328–338.

De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2011). Assessing collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education*, *14*, 201-206.

Dillenbourg, P. (2002). Over-scripting CSCL. In P. A. Kirschner (Ed.), *Three worlds of CSCL: Can we support CSCL?* (pp. 61-91). Heerlen: Open University of the Netherlands.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research in computer-supported collaborative learning: from design to orchestration. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-enhanced learning: Principles and products* (pp. 3-19). Springer.

Duijnhouwer, H., Prins, F. J., & Stokking, K. M. (2012). Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process, and performance. *Learning and Instruction, 22*, 171–184.

Dysthe, O. (2004). *The challenges of assessment in a new learning culture*. The 32nd International NERA/NFPF Conference, Reykjavik, Iceland.

Falchikov, N. (1995). Improving feedback to and from students. In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 157-166). London: Kogan Page.

Falchikov, N. (2003). Involving students in assessment. *Psychology Learning and Teaching, 3*, 102–108.

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a Script Theory of Guidance in Computer-Supported Collaborative Learning. *Educational psychologist, 48*, 56–66.

Gielen, M., & De Wever, B. (2012). Peer Assessment in a Wiki: Product Improvement, Students' Learning And Perception Regarding Peer Feedback. *Procedia - Social and Behavioral Sciences, 69*, 585–594.

Gielen, M., & De Wever, B. (2013). Structuring the Peer Assessment Proces: Impact on Feedback Quality. In N. Rummel, M. Kapur, N. Mitchell, & S. Puntambekar (Eds.), *To See the World and a Grain of Sand: Learning across Levels of Space, Time, and Scale: CSCL 2013 Conference Proceedings Volume 2 — Short Papers, Panels, Posters, Demos, & Community Events* (pp. 255-256). Wisconsin, MA: International Society of the Learning Sciences. Retrieved October 1, 2014, from http://www.isls.org/cscl2013/Volume%202%20Final%20CSCL%202013%20Proceedings.pdf

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*, 304–315.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*, 1509–1528.

Hattie, J. & Gan, M. (2011). Instruction based on feedback. In Mayer, R. E. & Alexander, P. (eds). *Handbook of research on learning and instruction* (pp. 249-271). NewYork: Routledge. Taylor and Francis Group.

Hattie, J., & Timperley, H. (2007). *The Power of Feedback*, *77*, 81–112.

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education, 71*, 133–152.

King, A. (2002). Structuring Peer Interaction to Promote High-Level Cognitive Processing. *Theory Into Practice, 41*, 33-39.

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

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*, 211-224.

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration scripts - A conceptual analysis. *Educational Psychology Review, 18*, 159–185.

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction, 20*, 344–348.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology, 41*, 525-536.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. Van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 125-143). Mahwah, NJ: Erlbaum.

Nelson, M. M., & Schunn, C. D. (2008). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.

O'Donnell, A. M. (1999). Structuring dyadic interaction through scripted cooperation. In: A. M. O'Donnell, & A. King (eds.), *Cognitive perspectives on peer learning*. (pp. 179–196). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Donnell, A. M., & Dansereau, D. F. (1992). Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In R. Hertz-Lazarowitz & N. Miller (Eds.), *Interaction in cooperative groups: The theoretical anatomy of group learning* (pp. 120–141). Cambridge, MA: Cambridge University Press.

Nicol, D., & Macfarlane-Dick, D. (2004). Rethinking formative assessment in HE: A theoretical model and seven principles of good feedback practice.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*, 199–218.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education, 39*, 102-122.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*, 195-203.

Poverjuc, O., Brooks, V., & Wray, D. (2012). Using peer feedback in a Master's programme: a multiple case study. *Teaching in Higher Education, 17*, 465–477.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11*, 289-303.

Sadler, D.R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment and Evaluation in Higher Education, 35*, 535–550.

Schellens, T., & Valcke, M. (2006). Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education, 46*, 349–370.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153 – 189.

Sluijsmans, D. (2002). *Student involvement in assessment: the training of peer assessment skills*, unpublished doctoral dissertation, Open University of the Netherlands, Heerlen.

Sluijsmans, D., Brand-Gruwel, S., & Van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education. *Assessment and Evaluation in Higher Education, 27*, 443–454.

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

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*, 291-303.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*, 265-269.

Strijbos, J. W., Van Goozen, B., & Prins, F. (2012, August). *Developing a coding scheme for analysing peer feedback messages*. Paper presented at the meeting of EARLI-SIG 1 Assessment and Evaluation Conference, Brussels, Belgium.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276.

Topping, K .J. (2009). Peer Assessment. *Theory Into Practice, 48*, 20–27.

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*, 1161-1174.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Berg, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*, 316-327.

Van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the Load Off a Learner's Mind : Instructional Design for Complex Learning. *Educational Psychologist, 38*, 5–13.

Tables

Table 1 – Scoring rubric to measure the quality of peer feedback messages

Main category	Sub category	Good Feedback		Average Feedback		Minimal Feedback	
Criteria	Content	Comments on all feedback aspects, in combination with the associated criteria	30	Comments on some feedback aspects, in combination with the associated criteria	15	None or minimal comments	0
	Clarification	Clarification of all comments on feedback aspects	20	Clarification of some comments on feedback aspects	10	None or minimal clarification of comments on feedback aspects	0
Feedback	Comments	Equilibrium between positive and negative comments	10	Mainly positive comments	5	Mainly negative comments	0
	Asked questions	Multiple questions which stimulate reflection	10	One question which stimulates reflection	5	No asked questions present	0
	Examples	Useful examples	5	Unclear examples	2	No examples present	0
	Suggestions	Useful and concrete suggestions for future improvement; Constructive advice	10	Vague and abstract suggestions for future improvement	5	No suggestions for future improvement present	0
Writing	Structure	Clear structure	5	Unclear structure	2	No structure	0
	Formulation	Short formulations	5	Mainly keywords	2	Only keywords	0
	Style	Written in first person throughout the whole feedback message	5	Occasionally written in first person	2	No use of first person	0
TOTAL			100				

Note. Adapted from the Feedback Quality Index (Prins, Sluijsmans, & Kirscher, 2006).

Table 2 – Scoring rubric to measure the quality of the product

Main category	Sub category	Good abstract		Average abstract		Poor abstract	
Situating the study	Intention / focus	The intention or focus of this study is specific and clearly explained in the first paragraph	10	The intention or focus of this study is rather vague	5	The intention or focus of this study is not described	0
	Problem statement	The context of the problem is clearly described	10	The context of the problem is not clearly described	5	There is no or a minimal description of the problem statement given	0
	Consistency	The intention or focus in combination with the problem statement forms a logical whole	10	The intention or focus of the study does not sufficiently reflect the problem	5	The intention or focus of the study and the problem are independent of each other	0
Content of the abstract	Methodology	The methodology is clearly explained and includes all details about setting	10	The methodology is rather vague and includes only limited details about setting	5	There is no or a minimal description of the methodology given	0
	Results	The main results of the study are concisely described and summarized	10	Not all the results of the study are described or too extensively discussed	5	The results are not summarized or are not addressed	0
	Limitations / suggestions for further research	The abstract refers briefly to the limitations of the studies and opportunities for future research	5	Little or no relevant limitations or suggestions are described in the abstract	2	No limitations or suggestions are described in the abstract	0
Finishing	Structure	Clear structure in line with the rules of an abstract	10	Unclear structure of the abstract	5	No structure	0
	Language	The abstract contains no grammatical errors and written in a smooth writing style	10	The abstract contains some grammatical errors. Insufficient attention was paid to the writing style	5	The abstract contains many grammatical errors. No attention paid to the writing style	0
	Word length	The length of the abstract corresponds to the agreed number of words	5	The length of the abstract is either just not long enough either slightly too long	2	The length of the abstract was not taken into account	0
General		The abstract shows that much time and attention was devoted to the task, leaving little or no modifications needed	20	The abstract shows that time and attention was devoted to the task, but there is still quite a lot of adjustments are necessary	10	The abstract shows that too little time and effort is spent on the task	0
TOTAL			**100**				

Table 3 – Multilevel models for the quality of the feedback (dependent variable: peer feedback score)

	Model 0	Model 1	Model 2 (final model)
Fixed			
Intercept (cons)	53.231(1.783)***	43.317(2.019)***	40.163(2.747)***
Time 2		13.119(1.640)***	13.119(1.640)***
Time 3		16.625(1.640)***	16.625(1.640)***
Basic structure			-1.717(3.601)
Elaborate structure			11.790(3.675)**
Random part			
Level 3 – Group	84.202(27.625)**	84.202(27.625)**	48.038(19.326)*
ρ(%)	19.32%	21.89%	13.77%
Level 2 - Student	48.961(20.086)*	74.561(19.406)***	74.864(19.442)***
ρ(%)	11.23%	19.38%	21.46%
Level 1 - Time	302.750(23.358)***	225.951(17.433)***	225.951(17.433)***
ρ(%)	69.45%	58.73%	64.77%
Model fit			
Deviance (-2LL)	4422.783	4324.474	4311.166
χ^2		98.309	13.308
df		2	2
p		$p<.001$	$p=.001$
Reference model		Model 0	Model 1

Notes: *$p < .05$. **$p < .01$. ***$p < .001$.

Table 4 - Multilevel models for the quality of the wiki product (dependent variable: product score)

	Model 0	Model 1	Model 2	Model 3 (final model)
Fixed				
Intercept (cons)	63.857(.860)***	39.329(1.111)***	38.452(1.637)***	41.059(1.814)***
Time 2		24.929 (0.993)***	24.929(0.993)***	22.439(1.684)***
Time 3		34.946 (0.993)***	34.946(0.993)***	29.614(1.684)***
Final version		9.139(0.811)***	9.139(0.811)***	9.139(0.801)***
Basic structure			1.637(2.077)	-0.699(2.488)
Elaborate structure			0.980(2.123)	-4.728(2.544)
Time 2 . Basic				1.165(2.372)
Time 3 . Basic				5.843(2.372)*
Time 2 . Elaborate				6.618(2.426)**
Time 3 . Elaborate				10.509(2.426)***
Random part				
Level 3 - Group	4.894(6.901)	4.894(6.901)	4.823(6.862)	4.823(6.862)
ρ(%)	1.02%	2.00%	1.97%	2.00%
Level 2 - Student	26.861(13.088)*	74.216(12.638)***	73.881(12.601)***	74.560(12.597)***
ρ(%)	5.58%	30.31%	30.22%	30.93%
Level 1 - Time	449.892(21.952)***	165.763(8.088)***	165.763(8.088)***	161.693(7.890)***
ρ(%)	93.40%	67.69%	67.81%	67.07%
Model fit				
Deviance (-2LL)	9077.192	8238.497	8237.871	8216.987
χ^2		838.695	0.626	20.884
df		3	2	4
p		***	*p*=.731	***
Reference model		Model 0	Model 1	Model 2

Notes: *$p < .05$. **$p < .01$. ***$p < .001$.