‘Now you know what you’re doing right and wrong!’ Peer Feedback Quality in Synchronous Peer Assessment in Secondary Education

Tijs Rotsaert[1], Ernesto Panadero[2], Tammy Schellens[1] & Annelies Raes[1]

Author note

[1] Department of Educational Studies, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

[2] Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Madrid, Spain

Recommended citation:
Rotsaert, T., Panadero, E., Schellens, T., & Raes, A. (2018). “Now you know what you’re doing right and wrong!” Peer feedback quality in synchronous peer assessment in secondary education. European Journal of Psychology of Education, 33(2), 255-275. doi:10.1007/s10212-017-0329-x

Correspondence concerning this manuscript should be addressed to: Tijs Rotsaert. Department of Educational Studies, Ghent University, Henri Dunantlaan 2, BE9000 Ghent, Belgium. E-mail: Tijs.Rotsaert@UGent.be

Funding:
Second author funded by the Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad) under the Ramón y Cajal program (Reference number: RYC-2013-13469).

Abstract

This study explores the effects of peer assessment (PA) practice on peer feedback (PF) quality of 11th grade secondary education students (N=36). The PA setting was synchronous: anonymous assessors gave immediate PF using Mobile Response Technology. The design was quasi-experimental (experimental vs. control condition) with two independent variables: (a) PA practice (10 times) and (b) helping assessees to filter out relevant information they received, which should influence PF quality in subsequent tasks in which they were assessors. PF content analysis showed that PA practice improved PF quality: messages contained more negative verifications, and informative and suggestive elaborations after the intervention. However, no effects were found of filtering out relevant information on PF quality. Moreover, students' perceived peer feedback skills improved which was in correspondence with their actual quality improvement over time. Additionally, the perceived usefulness of the received feedback was rated high by all participants.

Keywords: Peer feedback, peer assessment, assessment scaffolds, practice effect.

Introduction

Peer assessment (PA) is a peer-assisted, collaborative learning arrangement that includes students assessing their fellow students' performance by providing feedback, which could be quantitative in nature (i.e. grades or ratings across assessment criteria) and/or qualitative (written or oral comments) (Topping, 1998). The process of assessing and commenting on the strengths and weaknesses of peers' work can help familiarize the assessor with the evaluation criteria, and in this way help to develop knowledge on what constitutes good work, and what needs to be avoided (Yu & Sung, 2015). Numerous studies on PA have shown various benefits for the learning process, such as improved student motivation, improved conceptual understanding, communication skills and self-assessment skills (e.g. Falchikov & Goldfinch, 2000; Topping, 2009).

Peer Feedback (PF) is an important component of PA being the information that one student provides to a peer (e.g. Topping, 1998). Research emphasizes that students require practice and training to become skilled peer assessors and assessees, who provide and receive high-quality PF (Sluijsmans, 2002). Researchers and teachers are thus challenged to implement assessment activities in which students are prompted to provide frequent PF, resulting in a frequent enactment of the peer-assessor role (Tsivitanidou & Constantinou, 2016). Additionally, it has been argued that more research is needed to explore the quality of PF in PA and how it is perceived by students, because PF does not automatically lead to positive outcomes (Shute, 2008). Furthermore, more insight is needed into the kind of support students need in order to improve the quality of the feedback they produce (e.g. Tsivitanidou, Zacharia, & Hovardas, 2011). Two different types of scaffolds for PA have been explored: (a) guiding questions (e.g. helping the assessee to "filter out" the relevant information), and (b) establishing criteria and use

of rubrics. The aim of this study is twofold: exploring the effect of PA practice and PA scaffolds - especially "filter out"- on PF quality.

Training and practice as a prerequisite for a valuable peer assessment activity

The success of PA depends, to a great extent, on whether students are able to acquire critical assessment skills and provide valid judgement of their peers' work (Liu & Li, 2013) . Sluijsmans, Brand-Gruwel, van Merriënboer and Martens (2004) indicate that the general PA skill consists of three mains skills: (1) defining assessment criteria: thinking about what is required and referring to the products; (2) judging the performance of a peer: reflecting upon and identifying strengths and weaknesses; and (3) providing feedback for future learning. Previous research indicated that the development of the first and third assessment skill can be successfully trained (Sluijsmans, 2002; Van Zundert, Sluijsmans, & Van Merriënboer, 2010) and/or supported through assessment scaffolds (see further). However, the second assessment skill of being able to judge the performance of a peer requires multiple enactments in both the assessor-role and assessee-role. The more practice in these PA processes, the more likely students will develop the expertise for making sound PA judgements (Liu & Carless, 2006; Panadero, 2016). Additionally, PA practice enhances students' ability to produce higher work quality themselves (Sadler, 2010) as they will be better able to apply the internalised assessment criteria themselves. The importance of multiple enactment in PA is also acknowledged in Boud's view on sustainable assessment in which the capacity for students to make judgements of their own work is seen as essential to stimulate lifelong learning (Boud & Soler, 2015). Within this framework, PA offers students the opportunity to practice their evaluative judgements, which has simultaneously an impact on the peer assessors' self-regulating skills (Nicol, 2010).

In sum, previous research shows that practice and training are crucial for the development of PA skills (Sluijsmans, 2002). However, literature on how practice enhances PA skills is scarce and, furthermore, it has not been explored what are its effects on PF quality.

When is peer feedback of good quality?

The quality of PF is crucial as it is the basis for PA and provides a platform for engaging students in an interactive and elaborative feedback discourse as well as in taking ownership of their learning (Hattie & Gan, 2011). Additionally, PF has the advantage of bringing students in a situation in which they 'are on the same wavelength', making PF more understandable and useful for them (Topping, 2003). When writing feedback, students have more opportunities to engage in important cognitive activities such as critical thinking (i.e. deciding what contributes a good piece of work), planning, monitoring, and regulation (Lin, Liu, & Yuan, 2001). In essence, well-formulated feedback should provide an answer to three questions: 'Where am I going?' (feed up), 'How am I going?' (feedback) and 'Where to next?' (feed forward) (Hattie & Timperley, 2007). As stated in the recent work by Reinholz (2015) there are three broad categories of feedback: (1) process-focused, (2) product-focused and (3) person-focused feedback. Process-focused feedback encompasses both task feedback (i.e. whether or not the task is correctly completed) and self-regulation (i.e. how students monitor, self-control and direct their work during the task). Product-focused feedback relates to the (in)correctness of the task and why this is the case. Person-focused feedback is related to the person who is engaged in the task. Building on the recent work by Gielen and De Wever (2015) in this study we focus on product-focused feedback as we want explore how students are able to improve their PF skills through practice and support of guiding questions in a synchronous anonymous PA setting in which immediate feedback is given (see further). Previous research indicated that qualitative

feedback should provide two types of information: verifications and elaborations (Narciss, 2008). Verification refers to 'a dichotomous judgment to indicate that a response is right or wrong' and elaboration refers to 'relevant information to help the learner in error correction' (Hattie & Gan, 2011, p. 253). These types of information are thus seen as the structural components of feedback, because students require feedback that tells them not only if they performed the task correctly, but also why and what they should do about it to improve their work (e.g. Prins, Sluijsmans, Kirschner, & Strijbos, 2010). Therefore, offering elaborations that justify the verification (e.g. correct vs. incorrect) is presumed to be beneficial for students' learning. As a consequence a balanced proportion of verifications and elaborations is more valuable than just providing verifications alone (Gielen & De Wever, 2015).

Regarding the specific case of PF quality in PA settings, this has been explored in a series of recent studies by Gielen and De Wever (2012, 2015). These authors explored whether the use of guiding questions (e.g. "What would you change in your peer's work?") influenced PF quality. Regarding verifications, it was found that student usually tend to give mostly positive verifications (i.e. this is correct) in PA. However, the guiding questions used in the experimental conditions resulted in more negative verifications. According to the authors, this resulted into better PF because it provoked a more balanced proportion of positive and negative verifications, and therefore it was more descriptive of the actual performance rather than just pointing out to the positive aspects as PA assessors tend to do. Regarding elaboration Gielen and De Wever found that the guiding questions effect resulted into more suggestive elaborations (i.e. feedback on how to improve a future performance) but did not have an effect on informative elaborations (i.e. feedback on why a criterion was performed correctly or not). Finally, no differences were noted between the proportion of informative and suggestive elaborations.

Importantly, Gielen and De Wever explored the effects of the guiding questions in PF quality with higher education students, practicing 3 times PA and in an asynchronous (i.e. non-immediate PF) wiki environment. In the present study we wanted to check whether the effect of guiding questions in improving PF quality would remain with secondary education students, when there is a stronger PA practice (10 PA occasions over an school year) and organized in a synchronous PA setting (i.e. immediate PF). Originally, Gielen and De Wever used a setting where the assessors were non-anonymous. However, in the present study anonymity for the assessors is assured. The rationale behind making the assessors anonymous is to decrease negative effects as a consequence of interpersonal processes (Panadero, 2016; Vanderhoven, Raes, Montrieux, Rotsaert, & Schellens, 2015).

How to support students to provide high quality peer feedback in peer assessment settings?

PA is often described as a complex collaborative learning task that requires high-level cognitive processing (e.g. Kollar & Fischer, 2010). Therefore, any approach to help students to provide better PF to their classmates will have an impact on the PA implementation and, finally, on learning. Previous research has explored two initiatives to support students in providing high quality PF. First, offering guiding questions/guidelines on what good PF quality constitutes (Reinholz, 2015). The logic behind is that, by offering such questions, the students will reflect more about the PA exercise which thus becomes a more metacognitive activity. This type of questions can be used to help PA assessors in producing the feedback and/or by assessees to better understand the feedback received. There are multiple examples of these types of interventions such as the previously presented work by Gielen and De Wever (2012, 2015), who provided a template for the assessors. Another suggested approach has been to help assessees to filter out the feedback they received (Tsivitanidou & Constantinou, 2016). The hypothesis

behind this is that when assessees actively process the PF they receive, they will become better assessors in a subsequent task, which means that they will produce better quality PF. This hypothesis will be tested in the present study.

The second PA scaffold initiative involves the students in the selection of the PA assessment criteria and through the use of a rubric (Panadero et al., 2013; Sluijsmans, 2002). A rubric articulates the expectations for an assignment by listing the assessment criteria and by describing levels of quality in relation to each criterion (Reddy & Andrade, 2010). By using rubrics, students have a clearer understanding of what is expected of them as assessors and assessees, because rubrics provide assessment transparency (Panadero & Jonsson, 2013). A rubric is therefore often categorized as an assessment scaffold in PA research, one that has shown to increase the accuracy of PA (Panadero, Romero, & Strijbos, 2013). Therefore, rubrics are a positive support for PA. For that reason, all the participants in the present study will receive rubrics to enhance the potential of the PA tasks.

In sum, this study explores the effect of the first scaffolding approach, that is, the use of guiding questions by helping assessees to filter out the PA feedback. Two conditions will be compared: an experimental condition where students are actively supported to filter out PF vs. a control condition with no filter out scaffold. Additionally, both conditions help the assessors by providing them with guiding questions on how to assess and rubrics.

Importance of peer feedback skills and perceived usefulness

Identifying oneself as an active learner is a key element in the development of PF skills. For that reason, it is important to also incorporate students' perceived improvement of their feedback skills (Boud & Soler, 2015): if students perceive that they are becoming more capable as peer assessors, they will be more motivated to perform PA and believe it is useful. However,

this has not been explored in detail in previous research. Therefore, we will provide 10 PA occasions so that the participants will have plenty of experience with PA and so that we can explore the evolution of their perceptions along with their veridicality (because it is checked whether they were actually becoming a better PA assessor).

Furthermore, the willingness to follow the assessors' advice is essential to augment the quality of the performance (Boud, 2000; Nelson & Schunn, 2009). How students respond to PF is not just a feature of the activities themselves; this depends also on the ways in which PF is perceived useful (i.e. mindful reception of PF) which cannot be controlled in advance (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). Therefore, students' perceived usefulness of the received PF will also be explored in this study.

Research questions & hypotheses

The aim of this paper is twofold. First, to analyze the effect of PA practice on PF quality over time. And, second, to explore the effect of a scaffold which helps assessees to filter out the feedback they received into providing PF of better quality in subsequent tasks. Additionally, it was explored whether students' perceived an improvement of their PF skills, which will be compared to their actual evolution. Finally, we explored whether the PF is perceived as useful from an assessee's point of view.

Given these research aims, a PA setting in which participants could take the roles of both assessor and assessee was needed. Hence we created a reciprocal PA setting where groups of students assessed each others' work. Additionally, this study is organized around two performance and multiple PA cycles to explore the effects of practice (10 PA occasions). Consequently, this allows us to measure the evolution of PF quality over time. The specific research questions and hypotheses are:

RQ 1: What is the evolution in PF quality over time when students practice PA several times?

(H1) It is expected to find a practice effect with an increase of negative verifications and suggestive elaborations (Gielen & De Wever, 2012, 2015).

RQ2: What is the impact of helping assessees to filter out the feedback they receive on their own PF skills as assessors?

(H2a) Students in the experimental condition will provide PF of higher quality as assessors at an earlier occasion. Therefore, it is expected to find more negative verifications and suggestive elaborations in the experimental group. (H2b) Additionally, it is expected to find an interaction effect between the PA practice (RQ1) and the experimental condition (RQ2).

RQ3: Do the perceived PA skills change over time? Are they related to the actual change in the PF quality?

(H3) It is expected to find an increase of PA skills based on the PA practice (PA skills will increase over time) and the effect of the experimental manipulation (the participants in the experimental condition will perceive a faster increase of their skills).

RQ4: Did the students perceive the PF as useful?

(H4) It is expected to find an increase as students will have multiple occasions for PA practice. Due to the active scaffold of filtering out the received PF, it is expected that students in the experimental condition report a higher level of perceived usefulness.

Method

Participants

Participants in this study were 36 11th grade secondary students (MAge = 15, Range = 14-16) equally spread over two classes, with two different teachers. The majority was female (80.6%).

All students were enrolled in the theory-oriented general secondary education track and had no prior experience with the specific PA task (i.e. assessment of a group product).

Procedure

Students worked in small groups (12 groups) on a topic which they chose concerning a specific internship institute (e.g. a local library). The learning goal was to experience the valuableness of conducting research and explore the necessary skills. Therefore, the groups designed a research proposal and conducted it during their internship (e.g. analyzing costumers' buy intentions in a recycling store). In the first semester the groups presented the research proposal and, at the end of the second semester, they presented the results. Both presentations were done in front of the classroom group, and each presentation was assessed by their classmates. Assessors were told that their PF would not affect their course grade to avoid possible worries about the effects of PF.

The function of the PA activity was formative in nature as the teachers intended to let the students learn things from their peers' feedback. However, to stimulate effort and justify the investment of time in the presentation, the mean PA score of the group presentation was taken into account for 15% of the course grade.

Regarding the PA scaffolds, all the participants were involved in the selection of the assessment criteria: the teacher provided a rubric that was discussed in the classroom and changes were incorporated when needed. For example, for the presentation-related criterion *coherence between speakers* level one was proposed as *Speakers weren't aware about each others' content*. Students added a part they felt was missing: *Speakers weren't aware about each others' content, which resulted in the same content being told twice.* The final rubric (see Appendix 1) had six criterions, each with five quality levels. Out of the 6 criterions, three were

presentation-related, and the other three were content-related. Since the task was different from the first group presentation (introducing the research) and the second group presentation (presenting the results of the research) the three criterions changed. Additionally, the assessors in both conditions received three guidelines to support them while giving feedback: (1) make sure your feedback is specific and linked to the matching rubric criteria, (2) give suggestions for a future improved performance, and (3) appoint the strengths, but don't be hesitant to indicate weaknesses. Finally, the difference between the conditions was that the participants in the experimental group were asked to filter out the feedback they received via three guiding questions (see Appendix 2). As a means of a manipulation check, students in the experimental condition were asked to show and shortly elucidate the completed filter-out file to the teacher after they had received the FB of their first presentation.

Regarding the PA procedure, each student acted 10 times as an assessor within his/her class (i.e. 12 sessions per class) and 2 times as an assessee. As mentioned earlier, the identity of the assessor remained anonymous. This was facilitated through the use of Mobile Response Technology (MRT), in which assessors get the opportunity to give immediate anonymous PA scores (quantitative part) and PF (qualitative part) via web-enabled devices such as smartphones, tablets or laptops (Magaña & Marzano, 2014). In this study the free tool Socrative™ (Bèta Release) was used. Every PA session included three steps as depicted in Figure 1. After all the assessors evaluated the presenting group, the results were projected and verbally discussed in the classroom. The teacher moderated this discussion phase by asking reflective questions (e.g. what is the reason for the high number of remarks on the presentation structure?). Additionally, the Socrative reports (automatically generated anonymized Excel files) were sent to the assessed

group. It is important to mention that the teacher had the possibility to identify the assessors in case of unfriendly or hostile messages were given.

Insert **Fig. 1** PA-session

Measurements

The content analysis (RQ 1 & 2)

To measure the evolution of PF quality over time, the feedback content was analyzed at individual level at three occasions (henceforth FB Occasion 1, FB Occasion 2 and FB Occasion 3) on a subsample of 20 of a total of 24 sessions (6 sessions in December, 2 sessions beginning of June and 2 at the end of June per group). See Figure 2. This resulted in a database of 1561 segments.

Insert **Fig.2** Content analysis FB Occasions

The first two levels (i.e. PF style and PF type) of the hierarchical content-analysis scheme by Gielen and De Wever (2015) were used with a slight modification. PF style here includes two categories verification and elaboration; while the third category from Gielen and De Wever – general – was not included as it was not identified among our data. Regarding PF type, there are five categories: positive verification, neutral verification, negative verification, informative elaboration and suggestive elaboration (see Table 1). As only a very small amount of neutral verifications was found and only during FB Occasion 2, these were not presented in the analyses. Additionally, we added another level to our data: whether the PF related to a content-related criterion in the rubric or to a presentation-related criterion.

Insert Table 1.

Data were coded by the first author and an external coder that was trained for the task. A random subsample containing 38.41% of the total segments was coded by both with a Krippendorff's Alpha of .99 for the content-related criteria, and .98 for the presentation-related

criteria. Next, 7 out of the 20 feedback sessions were double coded resulting in 600 segments (267 content-related / 333 presentation-related).

The alpha values were above or equal to the popular benchmark of .80 (De Swert, 2012; Landis & Koch, 1977): content-related PF style (.88), presentation-related PF style (.98), verification type (.97) and elaboration type (.97).

Students' PF skills perception (RQ3)

Participants reported their PF capability using a 10-point slider scale (0 totally not capable – 10 totally capable; rounded to 1 decimal place), in 3 items (example item: *Rate your capability of being able to formulate suggestions for improvement regarding a peers' work*). This scale was measured before the start of the intervention ($\alpha = .79$), after the session in semester I ($\alpha = .88$), and after all sessions in semester II ($\alpha = .94$).

Students' perceived usefulness of the received PF (RQ4)

This variable was measured through a 3-items 7-points Likert-scale (example item: *The feedback in the Socrative report was useful for future presentations).* Reliability analysis showed acceptable scores ($\alpha^{\text{Semester I}} = .69$ / $\alpha^{\text{Semester II}} = .67$). Furthermore, an open-ended question was posed on this issue: *Do you think the feedback in the Socrative report was useful or not? Please, explain why (not).*

Data analysis

As mentioned earlier, the qualitative content data was treated quantitatively. Repeated measures ANOVAs were performed for all content categories with estimable amounts of feedback messages. The mean number of segments per assessor per session of a specific category was entered as an independent variable, and condition as between subjects variable.

Results

RQ1: What is the evolution in PF quality over time when students practice PA several times?

First, the results about verifications. In line with our hypothesis (H1) the effect of PA practice increased significantly the number of negative verifications in the presentation-related criteria category over time [$F(2, 68) = 2.68$, $p = .041$, $\eta_G^2 = .06$] (see Table 2). This means that after multiple sessions students gradually dare to indicate more weaknesses in a peer's work regarding presentation-related aspects. Additionally, contrast analyses revealed that between FB Occasion 1 and 3 [$F(1, 34) = 5.211$, $p = .029$, $r = .365$] there were significantly more negative verifications given. For the content-related negative verifications, there was no effect of time [$F(2, 68) = .41$, $p = .669$]. In relationship to the frequency of *positive verifications* there was a significant evolution: the number of content-related positive verifications changed among the different FB Occasion [$F(2, 68) = 9.48$, $p = .000$, $\eta_G^2 = .17$]: Occasion 1 and 2 [$F(1, 34) = 14.102$, $p = .001$, $r = .541$]; Occasion 1 and 3 [$F(1, 34) = 13.984$, $p = .001$, $r = .539$]. Regarding the presentation-related criteria no significant effect was found [$F(2, 68) = 2.326$, $p = .105$].

Insert Table 2

Regarding elaboration (Table 3), informative elaborations of content-related criteria increased [$F(2, 68) = 5.524$, $p = .006$, $\eta_G^2 = .115$]. Contrast analyses revealed a significant increase between FB Occasion 1 and FB Occasion 2 [$F(1, 34) = 4.155$, $p = .049$, $r = .329$] and FB Occasion 1 and FB Occasion 3 [$F(1, 34) = 10.019$, $p = .003$, $r = .477$]. Additionally for the informative elaborations in the PF messages on presentation-related criteria, a likewise increase was found: there was a significant main effect of Time [$F(1.55, 52.78) = 5.693$, $p = .01$, $\eta_G^2 = .108$] while applying a Greenhouse-Geisser correction. Contrast analyses revealed a significant difference between FB Occasion 1 and FB Occasion 2 [$F(1, 34) = 4.672$, $p = .038$, $r = .348$] and

FB Occasion 1 and FB Occasion 3 [F(1, 34) = 8.175, p =.007, r = .440]. Overall, we can say that students add more elaborative information in their PF messages when they get multiple practice opportunities.

As expected, there was a significant effect of practice in the suggestive elaborations in presentation-related PF messages, [F(2,68) = 5.875, p = .004, η_G^2 = .131]. Contrast analyses revealed a significant increase of suggestive elaborations between FB Occasion 1 and FB Occasion 3 [F(1, 34) = 12.982, p = .001, r = .524].. Regarding *suggestive elaborations* for the content-related criteria there was no significant main effect of practice [F(2,68) = 1.491, p = .232].

Insert Table 3.

In sum, it was expected that negative verifications and suggestive elaborations would increase as it was the case, but additionally informative elaborations also increased. This is actually an unexpected positive result as this adds to the previous evidence that through PA practice students improve the quality of the feedback they give as PA assessors.

RQ2: What is the impact of helping assessees to filter out the feedback they receive on their own PF skills as assessors?

Two hypotheses were tested here. First (H2a), it was explored whether the intervention in the experimental group (helping assessees to filter out information) would improve the PF quality they would give in subsequent tasks as assessors. This hypothesis has to be rejected as the intervention did not improve the PF quality in terms of negative verifications and suggestive elaborations, nor in other categories (see Table 4). Regarding H2b, it has to be rejected too as there was no effect in the interaction between practice and the experimental manipulation.

Insert Table 4

RQ 3: Students' PF skills perception

When assessing students' PF skills perception, during and after the PA-sessions, a repeated measures analysis indicates a significant main effect of practice [$F(2,70) = 7.64$, $p = .001$, $\eta_G^2 = .136$]. Contrast analyses revealed a significant increase between FB Occasion 1 and 2 [$F(1,35) = 10.32$, $p = .003$, $r = .477$] and FB Occasion 1 and 3 [$F(1,35) = 13.50$, $p = .001$, $r = .528$] (see Table 5). No significant effects where found for the interaction between practice and the experimental manipulation [$F(2,70) = .65$, $p = .53$] nor for the experimental manipulation [$F(1,35) = 2.50$, $p = .12$]. Therefore, the H3 can only be partially supported as there was an effect of practice but not of the experimental manipulation (i.e. filter out guiding questions).

Additionally, it was important to check whether the students' perceptions about PF skills increase was veridical. For that reason the concurrence between PF skills perception increase and the development of PF skills was explored. As students reported higher PF skills and the content analysis indicated an improvement of negative verifications, informative elaborations and suggestive elaborations over time, this suggests that students did not only improved their PF skills throughout the PA sessions as shown through content analysis, but that they were also aware of the learning progress they make regarding their feedback skills.

Insert table 5.

RQ 4: Students' perceived usefulness of PF

Students' perceived usefulness of the PF was highly positively evaluated after both their presentation in the first ($M_{Experimental} = 5.80$; $SD = .84$ / $M_{Control} = 5.70$; $SD = .62$) and second ($M_{Experimental} = 5.43$; $SD = .75$ / $M_{Control} = 5.33$; $SD = .81$) semester on a 7-point Likert-scale. Contrary to our H4 there is a significant decrease over time [$F(1, 33) = 5.36$, $p = .03$, $\eta_G^2 = .083$].

Despite the significant decrease, the absolute values on the 7-point Likert-scale remain highly positive. Additionally, no significant difference could be found between the two conditions [$F(1, 33) = 0.00$, $p = .99$].

Results from the open-ended question about the usefulness of the feedback in the Socrative report were very positive: 88.88% (semester I) and 94.44% (semester II) for the experimental condition and 100% (semester I) and 94.44% (semester II) for the control condition. Additionally, students stated that the plurality of opinions they received is the biggest advantage of this PA procedure.

Discussion

PF quality in PA settings is crucial for students to improve their work (Topping, 1998). Although previous research has shown that practice is central for the development of judgmental skills, literature on what exactly is the effect of PA practice on PF quality was not conducted yet. The main aim of this study was twofold: exploring the effect of both PA practice and PA scaffolds – especially "filter out" – on PF quality. Additionally, it was explored whether students' perceived an improvement of their PF skills, which was matched with their actual evolution. Finally, we explored whether the PF is perceived as useful from an assessees' point of view. The setting was a synchronous (i.e. immediate feedback) PA environment in secondary education where assessors were anonymous.

Regarding, our first hypothesis (H1) it can be maintained that PA practice has an impact on PF quality over time. Aligned with the research by Gielen and De Wever (2015), (a) our participants showed more negative verifications, and (b) more suggestive elaborations. However, next to that, our research is the first one to find an increase in informative elaborations. This significant increase might be explained by the fact that all assessors were scaffolded through a

combination of the rubric and the guiding questions. As informational feedback is beneficial for the students' performance, this is important because it means that at the end of our intervention, students were able to include the three utmost difficult feedback components in their immediate feedback messages: (1) mentioning what is not in line with the performance criteria, (2) why this is the case and (3) how it should be improved. The number of positive verifications remained stable, which confirms the finding of Gielen and De Wever (2015) that students usually tend to give positive verifications. This is not problematic, as long as the other PF components are also present. Overall, our findings confirm the importance of practice in PA settings since students' expertise for making valuable judgments on a peers' work improves over time (e.g. Boud & Soler, 2015; Panadero et al., 2016, Sluijsmans, 2002).

Although it was expected (H2a) that in the experimental condition (i.e. helping assessees to filter out useful feedback) assessors would provide PF of higher quality, the hypothesis needs to be rejected. Furthermore, hypothesis 2b cannot be maintained as no interaction effect between PA practice and the experimental condition was found. These non-significant effects might be explained by the fact that the effect of filtering out might not have been strong enough on its own and that its effect might have been diluted by the other PA scaffolds. For example, the fact that students in both conditions received a Socrative feedback report might have been a sufficient support to do an adequate feedback filtering, resulting in an increase of PF quality in the consecutive PA sessions regardless of the filter out effect. Likewise, the use of rubrics might have had a substantive positive effect on the quality improvement of the PF over time. As suggested by Cheng (2015), the use of rubrics itself could function as an incentive to make structured comments and may reinforce the PF quality. This leads to the conclusion that the

filter-out activity for the assessees could possibly be seen as an example of an over-scripting activity, at least within this specific context and for this specific task (Dillenbourg, 2002).

The findings regarding students' perceived evolution in PF skills (H3) are in line with the findings of their actual improvement over time. This finding suggests that all students acknowledge that their involvement in multiple PA sessions leads to an improvement of their PF skills. This is an important finding as identifying oneself as an active learner is considered to be a key element in the development of PF skills (Boud & Soler, 2015). Through providing and receiving PF multiple times, students are constructing meaningful feedback conceptions for themselves. Previous research has shown that this ensues several benefits such as giving students more control over the feedback processes, and as a result of this, also more control over their own learning (Nicol, Thomson, & Breslin, 2014). However, as students' perceived PF skill was measured through a more general quantitative scale compared to the detailed qualitative analysis of the actual FB, future research might also include a more general PF quality score (for example the Peer Feedback Quality Index by Prins, Sluijsmans and Kirschner (2006)) as both measures would then be comparable through a correlation of the actual gain and students' perceived gain over time.

The assessees' perceived usefulness of the received PF (RQ4) was also analyzed, as the willingness to follow assessors' advice is essential to augment the quality of the performance. As expected (H4), the results show that the received PF was highly positively appraised by students in both conditions. The small but significant decrease in appreciation of the PF over time might be related to the fact that students did not get the opportunity to tackle the suggestions for improvement after the second PA session in semester II, which was the case in semester I. Another reason, in line with the self-determination theory by Ryan and Deci (2000), may have

been the fact that students perceived the PF as something that controlled them, in a way that it hindered their self-perceived autonomy. Nevertheless, this should not be perceived as problematic since the overall score remained high. The fact that the PF was perceived useful is valuable as it is generally too easily assumed that students automatically perceive feedback as being useful (Harks, Rakoczy, Hattie, Besser, & Klieme, 2013). Furthermore, in a recent theoretical model on PA, sound feedback reception is seen as essential, because feedback helps students to form a more objective lens for self-assessment and self-regulatory processes (Reinholz, 2015). Therefore, we suggest that future research should study the impact of multiple enactments as peer assessors on students' self-regulating skills more thoroughly.

Given the sample size and gender bias of the sample (mostly female), the findings of this study should be interpreted with caution. Studies with bigger sample sizes, within other settings (e.g. higher education) and a variety of courses should be conducted in order to confirm the results of this study. In line with the recent work by Reinholz (2015) more sustained training might be needed to stimulate the feedback improvement even more. Furthermore, the impact of the PF received – although it was perceived useful - on the actual performance was not explored in this study, as advocated by Evans (2013). Future research should focus on this issue by defining quality categories for the task on which PA is performed, for example based on tasks of previous student cohorts and in close collaboration with the involved teachers. This would offer the opportunity to get insight into the impact of the different peer feedback styles.

Our findings are important for educational research and practice. Our study reveals that students in a PA setting improve the quality of their PF over time, and therefore practice should be a major component in PA implementations. As discussed by Panadero, Jonsson & Strijbos (2016) additionally to assuring such practice, teachers need to monitor the PA process and coach

the students even providing feedback about the PA itself. Another important implication of our results, is that PF quality is not only mentioning if something is correct or not (i.e. positive and negative verifications), but also offer information on why this was (in)correct (i.e. informative elaborations), in combination with suggestions to improve the presentations (i.e. suggestive elaborations). This is absolutely in line with previous research on how to give adequate feedback to promote learning (Hattie & Gan, 2011). Concerning possible practice constraints of implementing multiple PA sessions, MRT has proven to be adequate to facilitate the reciprocal feedback processes, so teachers are encouraged to use this tool to organize PA practice within their classrooms.

In conclusion, our study clearly shows that when students are offered PA practice opportunities in combination with rubrics and guiding questions for the assessors (not the assessees), the more likely students will develop expertise for making sound evaluative judgements on peers' work. More specifically, content analysis of the PF messages revealed that students not only inform their peers about what is wrong and why but also provide suggestions on how to improve the performance. In sum, this study clearly indicates that PA practice in combination with clearly defined assessment scaffolds constitutes a valuable classroom assessment practice since students experience a tangible educational value of PA through the perceived and actual growth of their PF skills.

References

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. doi:10.3102/00346543061002213

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. doi:10.1080/713695728

Boud, D., & Soler, R. (2015). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, 1–14. doi:10.1080/02602938.2015.1018133

Cheng, K.-H., Liang, J.-C., & Tsai, C.-C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education*, *25*, 78–84. doi:10.1016/j.iheduc.2015.02.001

De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. Retrieved from http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. Heerlen, Open Universiteit Nederland. Retrieved from https://telearn.archives-ouvertes.fr/hal-00190230/

Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, *83*, 70–120.

Falchikov, N., & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, *70*(3), 287–322. doi:10.3102/00346543070003287

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. doi:10.1016/j.sbspro.2012.11.450

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, *52*, 315–325. doi:10.1016/j.chb.2015.06.019

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2013). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology*, 1–22.

Hattie, J., & Gan, M. (2011). Instruction Based on Feedback. In *Handbook of Research on Learning and Instruction* (pp. 249–270). Routledge. doi:10.4324/9780203839089.ch13

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi:10.3102/003465430298487

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*, 344–348. doi:10.1016/j.learninstruc.2009.08.005

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. doi:http://doi.org/10.2307/2529310

Lin, S. s. j., Liu, E. z. f., & Yuan, S. m. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*(4), 420–432. doi:10.1046/j.0266-4909.2001.00198.x

Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, *11*, 279–290.

Liu, X., & Li, L. (2013). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education*, *39*(3), 1–18. doi:10.1080/02602938.2013.823540

Magaña, S., & Marzano, R. J. (2014). Using Polling Technologies to Close Feedback Gaps. *Educational Leadership*, 82–83.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J.M. Spector, M. D. Merril, J. van Merriënboer, M.P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology*. pp. 124–143. New York: Lawrence Erdbaum Associates.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, *37*, 375–401.

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501–517. doi:10.1080/02602931003786559

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122. doi:10.1080/02602938.2013.795518

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Human factors and social conditions of assessment. New York: Routledge* (pp. 1–39). New York, NY: Routledge.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144. doi:10.1016/j.edurev.2013.01.002

Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), Assessment for Learning: Meeting the challenge of implementation.

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*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005

Prins, F. J., Sluijsmans, D. M. A., & Kirschner, P.A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education*, *11*(3), 289–303.

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. (2010). Formative peer assessment in a CSCL environment: a case study. *Assessment & Evaluation in Higher Education*, *30*(4), 417–444. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02602930500099219

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*, 435–448. doi:10.1080/02602930902862859

Reinholz, D. (2015). Peer conferences in calculus: the impact of systematic training. *Assessment & Evaluation in Higher Education*, 1-17. http://dx.doi.org/10.1080/02602938.2015.1077197.

Reinholz, D. L. (2015). Peer-Assisted Reflection: A Design-Based Intervention for Improving Success in Calculus. International Journal of Research in Undergraduate Mathematics Education, 1(2), 234–267. doi:10.1007/s40753-015-0005-y

Reinholz, D. (2015). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, 1–15. doi:10.1080/02602938.2015.1008982

Ryan, R. M., & Deci, E. L. (2000). Self-determination Theory and the facilitation of intrinsic motivaton, social development and well-being. *American Psychologist*, *55*(1), 68–78. doi:10.1037/0003-066X.55.1.68

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

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

Sluijsmans, D. M. A. (2002). Student involvment in assessment: The training of peer assessment skills. (Unpublished doctoral dissertation). Open University of the Netherlands, Heerlen.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education and Teaching International*, *41*, 59–78. doi:10.1080/1470329032000172720

Topping, K. J. (1998). Peer Assessment between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249. doi:10.2307/1170598

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), (pp. 55–87). Dordrecht, The Netherlands: Kluwer Academic.

Topping, K. J. (2009). Peer Assessment. *Theory Into Practice*, *48*(1), 20–27. doi:10.1080/00405840802577569

Tsivitanidou, O. E., & Constantinou, C. P. (2016). A study of students' heuristics and strategy patterns in web-based reciprocal peer assessment for science learning. *The Internet and Higher Education*, *29*, 12–22. doi:10.1016/j.iheduc.2015.11.002

Tsivitanidou, O. E., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction*, *21*(4), 506–519. doi:10.1016/j.learninstruc.2010.08.002

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*, 270–279.

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education*, *81*, 123–132. doi:10.1016/j.compedu.2014.10.001

Yu, F.-Y., & Sung, S. (2015). A mixed methods approach to the assessor's targeting behavior during online peer assessment: effects of anonymity and underlying reasons. *Interactive Learning Environments*, 1–18. doi:10.1080/10494820.2015.1041405

Table 1. Coding scheme for analysing PF content quality (modification based on Gielen et al., 2015)

Category	Subcategory	Description	Examples
PF style	Verification	The feedback segment is an evaluative statement expressed as a positive, neutral or negative remark on past performance.	Content-related: *You gave a good explanation of which steps you're planning to take.* Presentation-related: *The switch between the speakers was smooth and without interruptions.*
	Elaboration	The feedback segment is an informative statement that builds further on verification or remark expressed as e.g. a question, a confirmation, a suggestion or a justification.	Content-related: *[...] because I think it will be difficult for such young children to fill up a questionnaire.* Presentation-related: *Try to look a bit more at the public.*
Verification type	Positive	The feedback segment is a positive evaluative statement.	Content-related: *The research methods were well chosen.* Presentation-related: *She speaks relaxed.*
	Negative	The feedback segment is a negative evaluative statement.	Content-related*: The connection between the research question and the conclusion was not that clear.* Presentation-related: *A lot of content was just read out loud.*
	Neutral	The feedback segment is a neutral evaluative statement.	Content-related: *It is a pity that your response rate was that low.* Presentation-related: *All dots were consecutively presented on a slide.*

Elaboration type	Informative	The feedback segment is an informative statement, which gives more details about a previous evaluative statement without activating the student to adapt his work.	Content-related*: Good that you used graphics (pos. verification),* ***that makes it more understandable.*** Presentation-related: *I thought the PPT was confusing (neg. verification),* ***maybe this was due to the particular theme.***
	Suggestive	The feedback segment is a suggestive statement, which gives more details about a previous evaluative statement with the purpose to activating the student to adapt his work.	Content-related: *For future presentations, try to give some more explanation on the categories in the diagrams*. Presentation-related: *Try to divide the texts fragments more amongst all participating speakers.*

Table 2. Verification type: descriptives, mean amount of positive and negative elaborations per student per session for content- and presentation-related criteria

Verification type		Occasion 1 M*(SD)*	Occasion 2 M*(SD)*	Occasion 3 M*(SD)*
Content-related criteria	Positive	.94*(.50)* [u]	.56*(.61)* [u,v]	1.10*(.78)* [v]
	Negative	.29*(.29)*	.24*(.37)*	.31*(.47)*
Presentation-related criteria	Positive	1.01*(.47)*	1.21*(.81)*	1.42*(1.05)*
	Negative	.37*(.27)* [w]	.45*(.54)*	*.63(.71)* [w]

Note: same superscript u,v,w indicate significant differences at $p < .05$

Table 3. Elaboration type: descriptives, mean amount of informative and suggestive elaborations per student per session for content- and presentation-related criteria

Elaboration type		Occasion 1 M*(SD)*	Occasion 2 M*(SD)*	Occasion 3 M*(SD)*
Content-related criteria	Informative	.17*(.20)*^{u,v}	.32*(.40)*^{u}	.47*(.56)*^{v}
	Suggestive	.14*(.18)*	.22*(.32)*	.14*(.28)*
Presentation-related criteria	Informative	.31*(.36)*^{w,x}	.53*(.63)*^{w}	.79*(1.01)*^{x}
	Suggestive	.16*(.19)*^{y}	.42*(.85)*	.79*(1.01)*^{y}

Note: same superscript u,v,w,x,y indicate significant differences at $p < .05$

Table 4. Test statistics for main and interaction effect RQ2

		Main effect	Interaction effect
Verification type			
Content-related criteria	Positive	$F(1, 34) = .546$, $p = .465$	$F(2, 68) = 3.133$, $p = .052$
	Negative	$F(1, 34) = .148$, $p = .703$	$F(2, 68) = 1.278$, $p = .285$
Presentation-related criteria	Positive	$F(1, 34) = .037$, $p = .849$	$F(2, 68) = 2.021$, $p = .141$
	Negative	$F(1, 34) = .002$, $p = .961$	$F(2, 68) = .633$, $p = .534$
Elaboration type			
Content-related criteria	Informative	$F(1, 34) = .224$, $p = .639$	$F(2, 68) = .577$, $p = .544$
	Suggestive	$F(1, 34) = .619$, $p = .437$	$F(2, 68) = .005$, $p = .995$
Presentation-related criteria	Informative	$F(1, 34) = .524$, $p = .472$	$F(1.55, 52.78) = 1.981$, $p = .157$
	Suggestive	$F(1, 34) = 1.693$, $p = .202$	$F(2, 68) = .749$, $p = .477$

Table 5. Descriptives students' PF skills perception

Control condition			**Experimental condition**		
Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
M*(SD)*	M*(SD)*	M*(SD)*	M*(SD)*	M*(SD)*	M*(SD)*
5.62*(1.35)*	6.24*(1.50)*	6.85*(1.23)*	6.19*(1.37)*	7.09*(1.67)*	7.07*(1.49)*

Fig. 1 PA Session

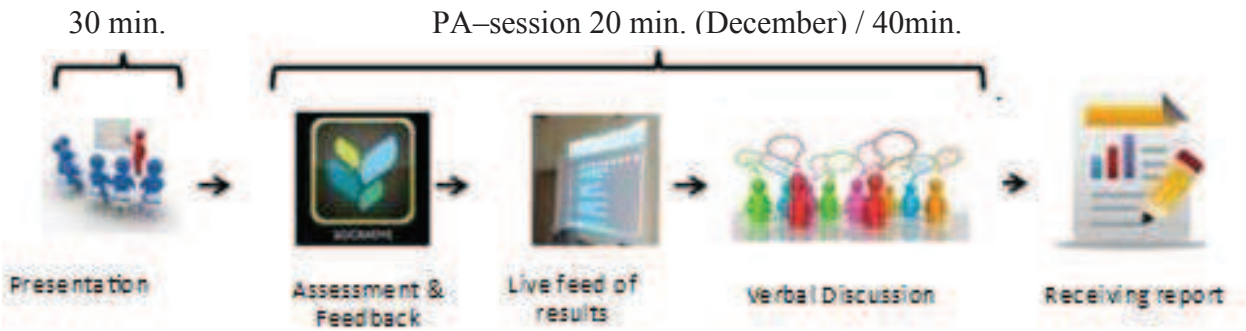

Fig.2 Content analysis FB Occasions