Taking a closer look at how higher education students process and use (discrepant) peer feedback

Accepted: version pre-print

Florence Van Meenen, Nicolas Masson, Leen Catrysse & Liesje Coertjens

florence.vanmeenen@uclouvain.be

Abstract:

Little is known on how students process peer feedback (PF) and use it to improve their work. We asked 59 participants to read the feedback of two peers on a fictional essay and to revise it, while we recorded their gaze behaviour. Regarding the PF processing subphase, discrepant PF led to more transitions, but only for participants who reported the discrepancy afterwards. Counterintuitively, participants who did not report the discrepancy, showed longer first-pass reading times. Concerning the PF use subphase, dwell time on essay correlated positively with the quality of the revised essays assessed by professors. Participants with a high-quality revision spent more time addressing higher order comments, corrected one or two lower order aspects at a time and proofread in the end, in which they went beyond the suggestions provided in the PF. These insights can be used when designing training to foster students' uptake of (discrepant) PF.

Keywords:

Peer feedback; feedback processing; eye movement; discrepancy

1. Introduction

In organisational sciences, feedback has frequently been labelled the food of the champions (Blanchard & Johnson, 2015). In educational sciences, feedback has been described a powerful lever for learning as well (Butler & Winne, 1995; Hattie, 2009). Feedback can come from different sources and recent meta-analyses suggest that peer feedback (PF) is highly effective. Indeed, Double, McGrane & Hopfenbeck (2020) summarised the effects of 54 (quasi-)experimental studies on peer assessment, 36 of which focused on written peer feedback. There was a significant positive small to medium effect of peer assessment, and more specifically of written peer feedback, on student learning. Wisniewski, Zierer and Hattie (2020) pooled the results from 435 studies, 8 of which contrasted teacher feedback with PF. Results indicate that PF fosters student learning more than teacher feedback. This suggests that, if feedback is the food of champions, PF may be a superfood.

Yet, how PF actually improves performance remains largely veiled (van der Kleij, 2020; van der Kleij & Lipnevich, 2020) Literature provides some ideas of what can hinder the uptake of teacher feedback

compared to PF. Two elements appear particularly salient. First, specific jargon or terminology has been highlighted as a key aspect of teacher feedback which could hamper the uptake of feedback (Garino, 2020; Hyland, 1998; Jonsson, 2013; Winstone, Nash, Parker, & Rowntree, 2017; Winstone, Mathlin, & Nash, 2019). PF may be easier to understand (Double et al., 2020), due to cognitive congruence between peers (i.e., "similarity in conceptual and factual knowledge", Altonji, Baños, & Harada, 2019, p. 446). Second, teacher feedback may be perceived as authoritative (Jonsson, 2013), leading students to accept it because it stems from the teacher (Hyland, 1998) and to limit their revision to the feedback provided (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010). Yet, reflection and the critical appraisal of PF (Huisman, Saab, van Driel, & van den Broek, 2018), which is incited due to social congruence (i.e., "closeness in authority or experience", Altonji et al., 2019, p. 446), may direct students to go beyond the aspects detailed in the feedback (Yang, Badger, & Yu, 2006). It should be noted that the vast majority of studies focus on writing skills (Falchikov & Goldfinch, 2000, Double et al., 2020). Therefore, differences in PF context depending on the field (i.e.: exact sciences, mathematic...) cannot be excluded (Double et al., 2020; Li, Xiong, Zang, Kornhaber & Lyu, 2016).

However, PF may be discrepant, which could hinder the uptake (Wichmann, Funk, & Rummel, 2018). Indeed, literature suggests including multiple peers, to increase the chances of receiving high-quality peer feedback (Huisman et al., 2018). Relying on multiple peers also appears to be the most frequent choice in practice settings (Cho & Schunn, 2018). This entails the risk of discrepancies (Aryadoust, 2016); one peer stating that a certain aspect of the task is adequate while another disagrees. Literature on multisource feedback in medical education suggests that most students find discrepant feedback disconcerting (Ginsburg, Vleuten, Eva, & Lingard, 2017), possibly leading them to ignore the feedback.

Despite the potential of PF for learning, studies focusing on the uptake of (peer) feedback remain scarce (Cutumisu, et al., 2019; Winstone, Nash, Parker, et al., 2017). Therefore, it is unclear how students handle discrepant PF. The present study examines how students process (discrepant) PF and which actions (i.e. changes that the learner makes after receiving the feedback, Van der Kleij & Lipnevich, 2020) lead to a high-quality revision.

1.1. Processing and use of feedback

Two key phases can be distinguished in FP: a first phase where students give feedback on their peers' work (Alqassab et al., 2018) and a second phase where students process and use the FPs received (Winstone et al., 2017). Different terms are used in literature to label the second phase that starts from the moment a student physically receives (peer) feedback and ends with the student handing in the

revised work[1]. Moreover, many authors described two distinct subphases in this phase of processing and use of PF(El Ebyary & Windeatt, 2019; Garino, 2020; Jonsson, 2013): a first that consists of reading and interpreting the feedback in relation to the first version of their work (which we will further label (peer) feedback processing) and a second in which students undertakes actions to improve his/her work based on the feedback provided (which we will label feedback use). Both subphases are considered indispensable for (peer) feedback to be effective (Buckingham & Aktuğ-Ekinci, 2017; Jonsson, 2013).

As the feedback processing subphase cannot be directly observed (El Ebyary & Windeatt, 2019), two online methods have been used: think aloud protocols and, in the last years, eye tracking (i.e., the capturing of a person's eye movements as an indication of a person's attention, Ashraf et al., 2018). The following sections will summarise the research findings from studies using either online method to examine processing and use of peer, teacher or computer-based feedback.

1.2. (Peer) feedback processing

In line with Richardson's (2013) recommendations, most studies on (peer) feedback processing use online measures; offline measures are generally poor indicators of task processing (Veenman, Bavelaar, De Wolf & Van Haaren, 2014) because they require a high capacity for reflexivity (Penttinen, Anto & Mikkilä-Erdmann, 2013) and only highlight conscious processes (Dimoka et al., 2013). Indeed, only the study by Garino (2019) relied on self-report. Five studies used online methods to better understand the first subphase in which students process the (peer) feedback. One study used think aloud protocols to examine processing of computer-based feedback (Máñez, Vidal-Abarca, Kendeou, & Martínez, 2019) and four studies used eye tracking: two studies examined computer-based feedback (Cutumisu et al., 2019; El Ebyary & Windeatt, 2019), while the other studies focused on PF (Berndt et al., 2018; Bolzer et al., 2015). In Berndt et al. (2018) and Bolzer et al. (2015), students read an essay of a fictional student and the feedback of fictional peers. This PF took the form of track changes comments in the study of Bolzer et al. (2015), while consisting of written criteria-based feedback in the study of Berndt et al. (2018).

The results of these studies suggest that students who focus longer on the PF and attempt to integrate it, recall it better afterwards. Indeed, Bolzer et al. (2015) detected a positive correlation between dwell

[1] proactive recipience of feedback (Winstone, Nash, Parker, et al., 2017; Winstone, Nash, Rowntree, & Parker, 2017; Winstone, Hepper, & Nash, 2021), feedback use (Kyaruzi, Strijbos, Ufer, & Brown, 2019), feedback engagement (Garino, 2020; van der Kleij, 2020), engagement with feedback (van der Kleij & Lipnevich, 2020), feedback processing (Cutumisu et al., 2019), feedback uptake (Wichmann et al., 2018), mindful reception (Gielen et al., 2010) and mindful cognitive processing (Berndt, Strijbos & Fischer, 2018; Bolzer, Strijbos, Fischer, 2015; Strijbos & Wichmann, 2018).

time on the PF and recall, though this was not confirmed in Berndt et al. (2018)'s study. In addition, Bolzer et al. (2015) found the number of transitions between the track changes comments and the essay, and between the different track changes remarks to be positively correlated with subsequent recall. This suggests that students recall more feedback elements when they have actively tried to integrate different feedback elements and the feedback in relation to the essay.

Results on the link between peer feedback processing and the quality of revision afterwards indicate, counterintuitively, that longer dwell times are associated with poorer revision. While Bolzer et al. (2015) was unable to find a significant relation between eye tracking measures and revision performance, Berndt et al. (2018) detected a negative correlation between dwell time on the criteria (which structured the PF) and revision performance. Longer dwell time could thus be indicative of difficulties in processing the feedback, which could lead to subsequent lower revision performance.

An alternative explanation for the counterintuitive finding may be that it is an artefact of the operationalisation of revision performance. The two studies which quantified such revision performance (Berndt et al., 2018; Bolzer et al., 2015) calculated it as the ratio of revisions (i.e., revisions made/total possible revisions), divided by the time spent revising. To the best of our knowledge, in educational practice, the revision time is seldomly considered and products are assessed in light of a pre-defined competence. Consequently, it remains to be examined whether the relation between dwell time and revision performance alters when using an operationalisation of revision performance that resembles higher educational practice more closely.

A limitation of the research on (peer) feedback processing is that the feedback was provided by only one sender (either a peer, teacher or computer). Therefore, the studies do not have any discrepant feedback elements. As a result, surprisingly, these studies did not investigate how learners deal with discrepant feedback. Hence, possibly, insights from literature on multimedia learning, examining text-picture discrepancies (Mudrick, Azevedo, & Taub, 2019; Schüler, 2017, 2019) and on text reading, examining intra-text (Braasch, Rouet, Vibert, & Britt, 2012; Hessel & Schroeder, 2020; Rayner, Chace, Slattery, & Ashby, 2006; Rinck, Gámez, Díaz, & De Vega, 2003; van Moort, Koornneef, & van den Broek, 2021) or inter-text discrepancies (Schoor, Rouet, & Britt, 2021; Stadtler, Scharrer, & Bromme, 2020) are insightful. These studies suggest that these discrepancies impact gaze behaviour in three ways. First, learners fixated more and spent more time processing discrepant information (Braasch et al., 2012; Hessel & Schroeder, 2020; Rayner et al., 2006; Schüler, 2017, 2019; Stadtler et al., 2020; van Moort et al., 2021). Second, learners confronted with discrepant information show more frequent re-reading of the conflicting information (Hessel & Schroeder, 2020; Rayner et al., 2006; Rinck et al., 2003;

Schüler, 2017; Stadtler et al., 2020; van Moort et al., 2021). Third, more transitions between discrepant parts are made: while Schüler (2019) did not detect an effect, other studies found learners to show more frequent transitions in order to reconcile the conflicting information (Mudrick et al., 2019; Rinck et al., 2003; Schoor et al., 2021; Schüler, 2017). Based on this literature, one could hypothesise learners to slow down when noticing a discrepancy in the PF, thus leading to longer first-pass reading times. To try to make sense of the discrepancy, they will look back (transitions) and re-read the discrepant feedback, leading to longer second-pass reading times.

1.3. (Peer) feedback use

Regarding the stage in which students act upon the (peer) feedback received, the scarce research evidence suggests that students follow the order in which feedback is presented and that feedback without a concrete suggestion, takes longer to act upon. These findings on the different actions undertaken stem from three recently published studies in the context of English as a foreign language, with either teacher (Ahmadian, Yazdani, & Mehri, 2019; Buckingham & Aktuğ-Ekinci, 2017) or computer-based feedback (El Ebyary & Windeatt, 2019). Students were found to adhere to the order in which feedback was presented (El Ebyary & Windeatt, 2019) and, when a concrete suggestion was lacking (for example, when feedback consisted of mere error flagging), students spent more time re-reading the first draft (Ahmadian et al., 2019). Empirical support is required on whether these first findings can be generalised to a PF setting.

1.4. The present study

This study sets out to examine how students process (discrepant) PF. In addition, this study investigates which revision actions are associated with a high-quality revision. Like previous studies (Falchikov & Goldfinch, 2000, Double et al., 2020), the present research also focuses on students' writing skills.

Regarding the *PF processing subphase*, we have a first, general research question, which leads us three hypotheses:

Research question 1: Are dwell time, number of transition and rereading related to PF recall and revision performance?

First, we will examine the relations as described by Bolzer et al. (2015):

H1: PF recall correlates positively with dwell time on the PF (H1a) and the number of transitions between the work and the PF and between the feedback of the different peers (H1b)

Second, the relation between dwell time and revision performance will be examined, using two operationalisations of such revision performance: (1) assessed by the researcher (Berndt et al., 2018;

Bolzer et al., 2015) and (2) assessed by professors in light of the competence without considering revision time, as to mimic higher education practice more closely. For the first operationalisation, the hypothesis made follows directly from the study conducted by Berndt et al. (2018):

H2: Dwell time on the PF correlates negatively with revision performance as assessed by the researcher (H2)

For the second, as no study has operationalised revision performance in this way, we describe the following research question:

Research question 2: Is there a relationship between dwell time and revision performance as assessed by professors?

PF entails the risk of discrepant views on the quality of the work. Based on research on discrepancies in multimedia learning and on text reading, the following hypothesis is formulated:

H3: Discrepant feedback is associated with longer dwell time during first-pass reading (H3a), longer second-pass reading (H3b) and more transitions between the feedback elements (H3c).

Regarding the *PF use subphase*, research evidence is scarce and none of the three previously published studies focused on PF; two of which focused on teacher feedback (Ahmadian et al., 2017) and the last on computer feedback (El Ebyary & Windeatt, 2019). To get insight into the actions associated with a high-quality revision, we will examine its' relation with gaze behaviour.

Research question 3: Is revision quality associated with dwell time or the number of transitions?

In addition, to better understand what sets participants with a high-quality revision apart, we will compare the revision action:

Research question 4: Do the revision actions and the order of these actions differ for participants with a low versus a high-quality revision?

2. Method

2.1. Ethical consent

We obtained ethical approval from the ethics review committee of the Faculty of Psychology at the XXX (file number EC XXXX) (blinded for peer review). Participants signed an informed consent prior to the experiment.

2.2. Participants and design

Sixty 2nd bachelor students (5 males and 55 females; *M age*=20.75, *SD*= 4.08) at the Faculty of Psychology (at university, country - blinded for peer review) voluntarily participated. All participants received course credits for their participation; which allowed them to complete a methodology course in their programme. Data from one participant had to be excluded due to technical issues.

As the third hypothesis required more detailed analyses, data from 16 participants were removed after visual inspection of the calibration accuracy (Holmqvist & Andersson, 2017). Thus, data from 43 participants (*M age*= 20.87, *SD*=4.57) were available for analyses on narrower areas of interest (AOIs).

Participants were asked to read an essay and the PF on this essay and to revise it (see Figure 1). A within-subjects design, in which participants received both discrepant and convergent feedback elements, was used to ensure that internal validity did not depend on the random assignment and to increase the statistical power (Charness, Gneezy & Kuhn, 2012).

[INSERT FIGURE 1 ABOUT HERE]

2.3. Materials and apparatus

2.3.1. Materials

The materials consisted of an instruction, assessment criteria, an essay of a fictional student and the feedback of two fictional peers (see Figure 2). The instruction and assessment criteria were created by the research team (see Appendix A). It described the context in which the fictional student had written the essay and the criteria that the fictional peers had used to provide feedback.

[INSERT FIGURE 2 ABOUT HERE]

The essay (281 words) was created by the research team and summarised the knowledge on the different types of memory, in line with the content taught in a compulsory 1st bachelor course. The feedback from two peers (56 words and 74 words, respectively), each containing five feedback elements, were created by the research team. To do so, we carefully examined the peer feedback provided in another course (in the 2nd bachelor in medical education), in which students wrote an essay and provided PF, which was found to be discrepant at times. We mimicked the student language and the discrepancy, while ensuring the content of the feedback matched the topic of the essay on types of memory. The feedback addressed six topics, in line with lower and higher order errors in the essay (Bouwer, Lesterhuis, Bonne, & De Maeyer, 2018): spelling mistakes (element 3), poor lay-out (convergent elements 1 and 6), the lack of a title (element 4), appraisal of the source credibility (convergent elements 2 and 8) and the lack of a definition (discrepant elements 5 and 10) and examples (element 9).

Two students at the Faculty of Psychology verified the essay and the PF and indicated that, in general, the quality resembled their peers' work. One aspect of the essay was too specific and one feedback element was too detailed according to the students. Adjustments were made in collaboration with the students. Subsequently, a pilot study was organised with 15 2nd bachelor psychology students, who confirmed that the essay and feedback resembled their peers' work. Yet, only one of the 15 students noticed the discrepancy in the peer feedback, as it was rather indirect (peer 1-element 5: the suggestion to add a definition on sensory memory vs. peer 2-element 10: "The topic is well summarised, concise and complete"). Consequently, the discrepancy was made more explicit (by adjusting peer 2-element 10: "No need to add definitions or examples of the different types of memory you mention").

2.3.2. Apparatus

The experiment was run on a PC equipped with a 22-inch LCD screen (1920x1080 pixels; refresh rate: 60Hz). An Eyelink 1000 infrared video-based pupil monitoring in a desktop remote mode camera was used to track monocular eye movements (SR Research, Mississauga, Canada; sampling rate: 500Hz; average accuracy range: 0.25° angle to 0.5° angle; gaze tracking range of 32° angle horizontally and 25° angle vertically). Participants were placed at approximatively 60 cm from the screen. The eye tracker was calibrated to the screen using a built-in 9-point protocol. The EyeLink Data Viewer software (SR Research, 2019) was employed for data extraction. This program allowed us to build videos of participants' fixations superimposed on the presented stimuli.

2.4. Measures

2.4.1. Gaze behaviour

To identify the fixations made by the participants, we used the EyeLink (SR Research Ltd., Ontario, Canada) default setting with a velocity threshold of 35 deg/s and an acceleration threshold of 8000 deg/s2 for fixation detection. To analyse learners' gaze behaviour, three areas of interest (AOIs) were created: feedback of peer 1 (AOI1), feedback of peer 2 (AOI2) and the essay (AOI3). In light of the hypothesis on the discrepant feedback, four additional areas of interest were created: around the discrepant (AOI4 and 5, dotted lines) and the convergent feedback elements (AOI6 and 7, continuous lines) (see Figure 2). These AOI's overlap with AOI1 and AOI2 but will not be used in the same analysis.

The following parameters were calculated to analyse gaze behaviour as a function of (discrepant) PF: First, dwell time on each of the AOIs was determined (Berndt et al., 2018; Bolzer et al., 2015; Cutumisu et al., 2019; Holmqvist & Andersson, 2017). Second, the number of transitions between AOI one to three, between AOIs four and five, and, between AOIs six and seven were computed. Third, for AOIs four and five and AOIs six and seven, dwell time for first-pass reading was extracted from the data

(Kaakinen & Hyönä, 2007). Second-pass reading was calculated (total dwell time - dwell time for first-pass reading). Since we aim to compare reading times between the feedback elements, first- and second-pass reading were normalized by calculating a milliseconds-per-character measure to account for the difference in length of feedback elements (Catrysse et al., 2018).

2.4.2. PF recall

In line with studies by Bolzer et al. (2015) and Berndt et al. (2018), PF recall was the ratio between the number of correctly recalled feedback elements for each of the two peers and the total number of feedback elements (being 10). This evaluation was carried out by the first author. A second researcher coded the PF recall for 33% of the participants. As inter-rater reliability was excellent (ICC=.905, p<.001) (Koo & Li, 2016), the first author coded the PF recall for the remaining participants.

2.4.3. Revision performance

Text revision performance was operationalised in two ways. Firstly, the number of feedback elements taken into account by the participant (further, number of PF elements used) was assessed by the first author. A second researcher assessed this as well for 33% of the participants. As the inter-rater reliability was substantial (Cohen's κ= .79, p<.001; McHugh, 2012), the first author assessed it for the remaining participants. Subsequently, in line with Bolzer et al. (2015) and Berndt et al. (2018), revision performance was assessed as the ratio of the number of PF elements used to the total number feedback topics (being 6, see Materials), divided by the time spent revising. For example, if a participant made revisions in line with 4 of the 6 PF elements and revised for 10 minutes, this would be 0.067 (i.e., (4/6)/10). We labelled this measure *Revision Performance Researcher*.

Second, 10 professors of the Faculty of Psychology (at university, country - blinded for peer review but the same faculty as the participants) appraised the quality of the essays (*Revision Performance Professors*), using comparative judgement method (in the Comproved tool, www.comproved.com). This method starts from the observation that people are more reliable in comparing two products than in assigning scores to those products. Hence, assessors are asked to compare two products and to indicate the better one, in light of a competence description (van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2019). Multiple assessors each make several comparisons, allowing to rank the products on an interval scale. This rank order represents the shared consensus of what constitutes a good product (Pollitt, 2012; van Daal et al., 2019). Comparative judgement has been found to provide valid and reliable appraisal of student work (van Daal et al., 2019; Verhavert, Bouwer, Donche, & De Maeyer, 2019) and its feasibility has been highlighted in contexts that do not require a grade to be fed back to students (Coertjens et al., 2021). We set out to obtain a .90 inter-rater reliability. Consequently, following guidelines from a recent meta-analysis (Verhavert et al., 2019), we aimed for 30 comparisons

per essay. To ensure a feasible workload for the raters, the 10 university professors were asked to complete 90 comparisons. The algorithm to select the pairs ensures distributed randomly drawn essays (for more detail, see Coertjens et al., 2021; Coenen et al., 2018).

Data from the 900 comparisons (wins-losses) were analysed using the Bradley-Terry-Luce model (Luce, 1959). Full detail on this model can be found in Coenen et al. (2018). The model generated a rank order of the essays from the weakest to the strongest essay (see Figure 3). For each essay, the analysis also provided a logit score for its quality, which ranged from –3.28 to 4.24. This logit score indicates the chance (more precisely, the logistic transformation of the chance) that an essay will win a comparison with an essay of average quality (having a logit score of 0).

[INSERT FIGURE 3 ABOUT HERE]

Subsequently, the reliability of the rank order was calculated using the Scale Separation reliability (Verhavert, 2019), which indicates to what degree the score distribution is not due to measurement error (Coenen et al., 2018): A minimal measurement error implies that the relative position of the items on the scale is quite fixed. The SSR for the rank order resulting from 900 comparisons was .89, indicating a high internal consistency of the rank order (Jones, Swan, & Pollitt, 2015).

2.4.4. Revision actions

Participants revised the essay on-screen, allowing the tracking of the gaze behaviour during this feedback use subphase (using the *EyeLink Data Viewer video)* and a screen capture (using CamStudio) of the mouse movements and the revisions made. Following the analysis strategy as detailed by Isohätälä, Näykki & Järvelä (2019), every 20 seconds we logged what the participant was modifying or reading.

We develop a coding scheme (see Table 1) following the work of El Ebyary et al. (2019); we established codes from the different items of the assessment criteria (see Material). Additional codes were added when participants made modifications that could not be directly linked to a feedback element. Subsequently, per participant, an overview was created of the sequence of their actions and the duration of these actions.

[INSERT TABLE 1 ABOUT HERE]

2.4.5. Manipulation check

In line with Rinck et al. (2002), the manipulation check was conducted during the interview at the end of the experiment (see Procedure). Participants were asked "Have you noticed any discrepancies between the feedback from the two peers?" (see Appendix B). If a participant had noted the discrepant feedback, they were asked to describe it in more detail.

2.5. Procedure

First, the participant was asked to read the instructions for the essay and the criteria (see Appendix A). The participant was informed that he/she would improve the essay of a student based on feedback from two peers. Subsequently, the participant was asked to place his or her head on a chin-rest in front of the eye-tracking device and an initial calibration was performed. After a satisfactory calibration, the recording of the first subphase (peer feedback processing) was launched. The participant was invited to read the essay and feedback that were displayed on the screen until he/she felt ready to modify the essay and the recording was stopped. Prior to the second subphase (PF use), the eye-tracker was anew calibrated. The participant then revised the essay. No time limit was imposed on the participant. Afterwards, the participant was asked to perform a distraction task for 10 minutes (Sudoku, Spot the difference, and Word Search) and was then asked to recall, with as much detail as possible, the PF. Last, the semi-directive interview gave participants the opportunity to express how they had improved the essay and if there was any feedback that they found more difficult to implement (see Appendix B).

The experiment lasted approximately 60 minutes, but it varied depending on the time needed for revising the essay.

2.6. Data analysis

To analyse the first two hypotheses and the second research question (i.e., H1: PF recall correlates positively with dwell time on the PF (H1a) and the number of transitions between the work and the PF and between the feedback of the different peers (H1b), H2: Dwell time on the PF correlates negatively with revision performance as assessed by the researcher, Research question 2: Is there a relationship between dwell time and revision performance as assessed by professors?), correlations were used on the data of 59 participants. Correlations exceeding .10, .30 and .50 were interpreted as small, medium and large, respectively (Cohen, 1988).

To verify if discrepant feedback is associated with longer dwell time during first-pass reading (H3a), longer second-pass reading (H3b) and more transition between the feedback elements (H3c), four additional AOIs were created (see Gaze behaviour in Measures). The data allowed this for 43 of the 59 participants (see Participants and design). Of these 43 participants, nine had reported the

contradiction in the manipulation check (further labelled Reporters), while 34 did not (Non-reporters). A repeated measures ANOVA was conducted with Feedback Discrepancy (discrepant – AOI 4 & 5 vs. convergent – AOI 6 & 7, see Figure 2) as a within-subject variable and Reporting Group (Reporters vs. Non-reporters) as a between-subject variable on the number of transitions made within the pairs of feedback elements (H3c) on the one hand and dwell time during first-pass (H3a) and second-pass reading (H3b) on the other hand. We used partial eta squared to estimate the effect size. Conventional effect size interpretations for partial eta squared suggested by Cohen (1988) are .0099 (small), 0.588 (medium) and .1370 (large). Paired-sample t-tests were performed when the repeated measures ANOVA showed a significant interaction. For these follow-up t-tests, Cohen's d effect size was calculated and values exceeding .2, .5 and .8 were interpreted small, medium and large, respectively (Cohen, 1988).

To analyse if revision quality is associated with dwell time or the number of transitions (research question 3), correlations were used. To shed light on whether the revision actions and the sequence of these actions differed for participants with low- versus high-quality revision (research question 2), we selected the participants with the 10 best revised essays (further labelled, high-quality revision) and the 10 weakest revised essays (further labelled, low-quality revision) according to the revision performance professors measure. In line with the approach taken by Van Gog, Paas and Merriënboer (2005) and Máñez et al. (2019), the separation of the two groups was done in order to maximize the contrast between them. It is based on a visual inspection of the rank order of the revision performance (see Figure 3). For each of the participants, the revision actions were analysed (mean duration video: 20 min. 57s. for the 10 high-quality revisions; 16 min 58s. for the 10 low-quality revisions). Subsequently, the overview of the revision actions of the participants with a high-quality revision were then compared to those with a low-quality revision. A independent samples t-test was performed to determine the differences in time allocated to the consideration of higher order comments between the two groups. Chi² tests were also performed. These allow us to determine whether the differences in terms of the revision actions carried out by the two groups of students were significant. Cramer's V effect size was calculated and values exceeding .1, .3 and .5 were interpreted small, medium and large, respectively (Kim, 2017).

All quantitative analyses were conducted using SPSS statistics (version 27).

3. Results

3.1. Feedback processing

Regarding the first hypothesis, the correlation was calculated between the dwell time on the PF (AOIs 1 + 2) during the processing subphase and PF recall (H1a) *(r*=.117; *p*=.378). In addition, the correlation between PF recall and two transition measures was calculated: the transition between the feedback from the two peers (AOI 1 vs. AOI2) on the one hand and between the essay (AOI3) and the PF (AOIs 1 + 2) on the other hand (H1b). Hypothesis 1 was not confirmed: none of these correlations reached the significance level (*r*=0.1; *p*=.450 and *r*=.05; *p*=.705 respectively).

Concerning the second hypothesis and the research question, the correlation between the dwell time on the PF (AOI 1 + 2) during the processing subphase and the revision performance as assessed by the researcher was examined (H2). The correlation between the dwell time on the PF (AOI 1 + 2) as assessed by the professors was performed (research question 2). Neither correlation was significant (*r*=-.045; *p*=.736 and *r*=-.208; *p*=.115, respectively) and hypothesis 2 was not confirmed.

Results partially confirm the association between discrepant feedback et le longer dwell time during first-pass reading (H3a), longer second-pass reading (H3b) and more transition between the feedback elements (H3c). The ANOVA on first-pass reading (Hypothesis 3a) indicated no main effect of Feedback Discrepancy (*F*(1, 41)=.006, *p*=.936) nor of Reporting Group (Reporters *vs.* Non-reporters, *F*(1, 41)=2.994, *p*=.091). However, there was a significant interaction (*F*(1, 41)=7.454, *p*=.009, η^2_p=.154): for the Reporters no difference was observed (*t*(8)=1.939, *p*=.089), while the Non-reporters read the discrepant feedback elements for longer on the first-pass reading (*M*=29.332 ms, *SD*=19.043) than the convergent elements (*M*=45.782 ms, *SD*=29.042; *t*(33)=-2.930, *p*=.006, Cohen's *d*=.502).

Regarding hypothesis 3b, no main effect of Feedback Discrepancy (*F*(1, 41)=0.570, *p*=.455) on second-pass reading was observed, but there was a main effect of Reporting Group (*F*(1, 41)=4.565, *p*=.039, η^2_p=.100): Reporters had longer second-pass reading times (*M*=75.831 ms, *SD*=11.265) than the Non-reporters (*M*=48.766 ms, *SD*=5.796). The interaction failed to reach significance (*F*(1, 41)=3.022, *p*=.090).

Concerning hypothesis 3c, results indicate a main effect of Feedback Discrepancy (*F*(1, 41)=19.482, *p*<.001, η^2_p =.32), indicating that participants made more transitions between the discrepant feedback elements (*M*=0.302, *SD*=0.513) than between the convergent feedback elements (*M*=0.116, *SD*=0.391). As Figure 4 shows, there was no main effect of Reporting Group (*F*(1, 41)=1.208, *p*=.278), but there was a significant interaction (*F*(1, 41)=13.676, *p*<.001, η^2_p=.25). While no significant difference was observed for the Non-reporters (*t*(33)=-0.812, *p*=.420), Reporters made more transitions between the discrepant feedback elements than the convergent feedback elements (*t*(8)=4, *p*=.004, Cohen's *d*=1.333).

[INSERT FIGURE 4 ABOUT HERE]

3.2. Feedback use

The number of transitions was not associated with revision quality (correlation revision performance professors: r=.055, p=.681; researcher r=-.146, p=.271) (research question 2). However, the total dwell time on the essay and on the feedback (AOIs 1-3) during the use subphases did show a significant positive correlation with Revision performance professors (r=.300, p=.021), while a significant negative correlation was detected with the Revision performance researcher (r=-.521, p<.001). Further analysis showed that these significant effects are due to differences regarding the dwell time on the essay (correlation Revision performance professors: r=.306, p=.018; researcher r=-.522, p<.001).

Regarding the revision actions used by students with a low versus high quality revision (research question 4), Table 2 presents an overview of the different actions and their sequence for one of the 10 participants with a high-quality and one of the 10 participants with a low-quality revision. Analysis of the (sequence of) actions for both groups of 10 participants revealed three differences[2].

[INSERT TABLE 2 ABOUT HERE]

First, participants with a low-quality revision addressed higher order comments to a lesser extent (Mean $_{\text{participants low-quality revision}}$=1 minute 19 seconds; Mean $_{\text{participants high-quality revision}}$=6 minutes 2 seconds; $t_{(18)}$=-7.435, p<.001, Cohen's d=-3.325). Of the ten participants with a low-quality revision, five did not address the higher order feedback elements. The other five only made very limited revisions, with four of them addressing only one of the three higher order feedback elements.

Second, all participants, except one, started the revision by modifying lower order aspects (adding a title, correcting spelling errors or creating paragraphs). Yet, the way to do this differed between participants with a low- or high-quality revision ($\chi^2_{(1)}$=12.8, p<.001, Cramer's V=0.8). The former showed a paragraph by paragraph approach: for each paragraph they made modifications based on multiple feedback elements such as spelling, punctuation and lay-out. The latter participants showed an entire essay approach: they modified an aspect in the entire essay before moving on to another lower order aspect (e.g., correction of spelling throughout the essay and then tackling punctuation). The top five participants modified one lower order aspects at a time, while the other five were found to combine multiple elements (e.g., correcting spelling and lay-out while reading once through the essay).

Third, the participants with a high-quality revision included a proofreading stage at the end, while participants with a low-quality revision did not ($\chi^2_{(1)}$=10.769, p=.001, Cramer's V=0.734). During this proofreading, the gaze behaviour indicated that participants fixated on the essay, without looking back

[2] The qualitative and quantitative analyses conducted on 5 vs. 5 participants and 15 vs. 15 participants showed similar results.

at the feedback. participants. In this proofreading, participants addressed the lower order aspects once again but went beyond the PF: they put keywords in bold, added connecting words and numbers or bullets to structure the essay, deleted unnecessary abbreviations and replaced nouns by pronouns to avoid repetition.

4. Discussion

This study examined how (discrepant) PF is processed and looked into the associations between revision actions and revision quality. Regarding the PF processing subphase, Hypothesis 1 was not confirmed: PF recall did not correlate with dwell time on the PF during the processing subphase (H1a) nor with the number of transitions between the essay and the PF and between the PF of the two peers during this subphase (H1b), which contradicts previous findings by Bolzer et al.'s (2015). Yet, Berndt et al. (2018) were not able to replicate the correlation between dwell time on the PF and PF recall either. It is plausible that participants' PF recall is also influenced by how well they engage with the PF when they revise (PF use subphase). This raises the question whether using PF recall is confusing means with ends: if the goal of PF is that a student hands in revised work of higher quality, keeping the PF in mind could help the student during the use subphase, yet, in most practice settings, the PF can be consulted anew. Consequently, future studies may consider using the PF recall as criterion for inclusion of a participant's data in the analyses (for a similar practice, see Rinck et al., 2003): compared to participants who rushed through the experiment, participants who actively engaged with the PF during the processing and use subphases, likely recall part of it afterwards.

In addition, there was no relation between the dwell time on the PF during the processing subphase and the Revision performance as assessed by the researcher (H2) or by the professors (research question 2). This finding contradicts previous findings by Berndt et al. (2018) who found a negative association between processing time and Revision performance, but confirms the findings by Bolzer et al. (2015). Possibly, the absence of specific jargon in the PF (Garino, 2020; Jonsson, 2013; Winstone, Nash, Parker, et al., 2017) and the authentic language of the PF, as confirmed in the pilot study, made it easy to understand (Double et al., 2020). Consequently, the average processing time was short (Mean time for the processing subphase: 3 mins 40s) and there was little variance in this processing time (range: 1 min 19s – 7 mins 52s). The PF processing subphase may thus have been too short and too similar to detect associations with revision quality.

Hence, the main reasons for the observed differences in the quality of the revised essays would be expected to occur during the PF use subphase. This study innovates in tracking participants gaze behaviour during this use subphase. No association was observed between the number of transitions

and the revision quality, but results indicate a relation with the dwell time on the essay (research question 3): longer dwell times were associated with lower revision quality as assessed by the researcher (large effect) and higher revision quality as assessed by the professors (medium effect). This shows the impact of the operationalisation of revision performance. In line with previous research (Bolzer et al., 2015; Berndt et al., 2018), the former is calculated as the proportion of feedback elements used divided by the time spent revising. As observed in the present study, and in line with findings in other studies (Aben, Timmermans, Dingyloudi, & Strijbos, 2021; Bouwer & Dirkx, 2021), some students only address lower order comments. These revisions can often be made rapidly (i.e., a small denominator), leading to high scores on revision performance as assessed by the researcher. We argue that, to resemble higher education practice more closely, the revision time should not be considered and assessment should be in light of the pre-defined competence, as was the case in the revision quality as assessed by the professors.

The results on the qualitative analysis (research question 4) corroborate the positive association between dwell times on the essay and revision quality, which aligns with findings by Bouwer and Dirkx (2021). First, participants with a high-quality revision spent significantly more time on the essay as they were addressing higher order aspects, while their colleagues with a low-quality revision left some or all higher order aspects unattended. To design interventions fostering PF uptake, further qualitative research is warranted on the reasons for leaving higher order feedback elements unattended. Second, participants with a high-quality revision tackled one or two lower order aspects throughout the entire essay before modifying another aspect. This difference in the revision actions by the two groups of students is significant. Previous research found an association between self-reported feedback use and conscientiousness (Winstone et al., 2021). Possibly, striving to use all feedback elements and tackling lower order aspects one-by-one, is related to this personality dimension. Future research should consider assessing conscientiousness in addition to online measures of feedback use. A last aspect that sets participants with a high-quality revision apart was the proofreading stage at the end, in which participants went beyond the feedback of their peers. This difference is significant as well. This "going the extra mile" was not detected in previous studies on feedback use, possibly due to the focus on teacher (Ahmadian et al., 2019; Buckingham & Aktuğ-Ekinci, 2017) and computer-based feedback (El Ebyary & Windeatt, 2019) in these studies, which may be perceived as authoritative by students (Gielen et al., 2010; Jonsson, 2013). This resonates with Yang et al.'s (2006) findings that, compared to teacher feedback, PF incited students to go beyond aspects detailed in the feedback, which can be linked to the social congruence aspect of PF (Altonji et al., 2019). Clearly, future experimental research is needed to verify whether students are more prone to going the extra mile in PF contexts. Such

research could anew include the personality dimension conscientiousness and explore the links with students' achievement goal orientations (Winstone et al., 2021; Fong et al., 2021).

It is noteworthy that regardless of the revision quality, participants did not appear to re-read the essay prior to dealing with certain (higher order) PF elements, thus contradicting findings by Ahmadian et al. (2019). A possible explanation is that the PF in this study ensured concrete suggestions for all elements, in contrast to the teacher feedback used in Ahmadian and colleagues (2019) work. An additional finding is that all participants started with one or multiple easy to address lower order aspects and that they did not adhere to the order in which the PF was presented, thus contradicting previous findings by El Ebyary and Windeatt (2019). Possibly, the focus on PF, in which some feedback elements of the peers converged, incited participants to choose the order in which they modified aspects. An alternative explanation is that in El Ebyary and Windeatt (2019)'s study, the computer-based feedback appeared to be ordered from easy ('grammar', 'usage') to difficult ('Organisation and development', which contained broader suggestions, requiring students to self-assess). Future research could deliberately put a higher order aspect as the first feedback element, while varying the feedback source (computer, teacher, peer).

The findings in research on multimedia learning (text-picture discrepancies) and reading research (intra and inter-text discrepancies) appear to be partially generalisable to discrepant PF (hypothesis 3). For the Non-reporters, discrepant feedback provoked longer first-pass reading (medium effect, confirming H3a, in line with Braasch et al., 2012; Hessel & Schroeder, 2020; Rayner et al., 2006; Schüler, 2017, 2019; Stadtler et al., 2020; van Moort et al., 2021). This appears counterintuitive: the gaze behaviour of the participants who failed to report the discrepancy afterwards, shows that they did notice the discrepancy. Possibly, the findings can be interpreted in light of Stadtler et al. 's (2020) stages of processing conflicting information in multiple documents. The conflict detection (first stage) concerns the noticing of the discrepancy, making readers slow down (Braasch et al., 2012). During the second stage, the conflict regulation, different processes can occur. Some participants may ignore the conflict, and thus fail to report it afterwards. This is a common finding in research on discrepancies (Schüler, 2017; Stadtler et al., 2020; Stadtler, Scharrer, Brummernhenrich, & Bromme, 2013): even with quite obvious contradictions (Rinck et al., 2003), a large proportion (64%, Rinck et al., 2003; 78%, Schüler, 2017) of participants did not report it afterwards, which aligns with the 79% Non-reporters in the present study.

An alternative way of conflict regulation is attempting to restore coherence by looking back and rereading (Braasch et al., 2012; Stadtler et al., 2020). This study provides partial empirical support for this. The Reporters made significantly more transitions between the Discrepant (compared to

Convergent) feedback elements (confirming H3c, large effect), which corroborates previous findings (Mudrick et al., 2019; Rinck et al., 2003; Schoor et al., 2021; Schüler, 2017). Rinck et al. (2003) for example detected that, compared to non-reporters, reporters more often looked back to a previously read sentence (of the discrepant duo of sentences). However, discrepant PF was not reread more than convergent PF (H3b), which is at odds with previous research on discrepancies (text-picture: Schüler, 2017; intra/inter text: Hessel & Schroeder, 2020; Rayner et al., 2006; Rinck et al., 2003; Stadtler et al., 2020; van Moort et al., 2021). While Reporters had significantly longer second-pass reading times than Non-reporters, the interaction failed to reach significance (p=.090). As the present study is the first to examine the processing of discrepant PF, replication research is obviously required to determine whether, on the aspect of second-pass reading, discrepant PF deviated from other discrepancies or whether the finding is due to statistical power in the present study.

To resolve the conflict stemming from discrepant information, previous research found that readers gather more information on the sources, to help them decide whom to believe (i.e., discrepancy-induced source comprehension assumption, Schoor et al., 2021). In the field of PF use, the perception of a peers' abilities has been found to impact the degree of feedback use as well (Aben et al. 2021). In the present study, however, the PF was anonymous and the feedback provided by the two peers was of similar quality, thus complicating this action for conflict resolution. Further research using think aloud or cued retrospective reporting could shed light on the degree to which and how students attempt to assess source expertise when PF is provided anonymously. Possibly, they rely on the language used by the peers to deduct their expertise level.

4.1. Limitations

Some limitations of the present study need to be acknowledged. First, to shed light on the variability within the group of Reporter and Non-reporters, the conflict verification task (developed by Stadtler et al., 2013) may be useful in future research. It allows for a more fine-grained way of evaluating participants' perception of the discrepancy, by correcting the proportion of identified discrepancies for acquiescence bias (Schoor et al., 2021; Stadtler et al., 2020). Second, in line with previous research (Berndt et al., 2018; Bolzer et al., 2015), participants improved an essay of a fictional peer.

The use of a fictional essay was essential for this first study; it allowed to observe similarities and differences in the way students process discrepant feedback elements. Nevertheless, given the findings in this study, future research is needed to discern which findings generalize to the processing and use of PF on students' own work. Indeed, possibly, dwell times on the essay are shorter when participants revise their own essay for example. For such research, it appears key to also consider students' self-efficacy (Prilop, Weber, Prins, & Kleindrecht, 2021) and emotions (Garino, 2020;

Lipnevich, Murano, Krannich, & Goetz, 2021) as well. However, quantitative analyses of gaze behaviour become much more complex when it concerns students' own work. Therefore, considering the limited set of studies using online methods (Jonsson, 2013; van der Kleij & Lipnevich, 2020), it also appears sensible to further gather knowledge on students' PF processing and use when improving the work of a fictional peer. Third, the order of the feedback of both peers (AOI 1 and AOI2) was constant in the present study. To reduce the effect of the order of presentation, a counterbalanced design could be useful in future studies (Atkinson et al., 2010): the feedback from peer 1 and peer 2 could be reversed for half the participants. However, as the dwell time for the feedback from peer 1 and peer 2 was summed in the present study, we believe that the effect of order of presentation did not influence the results. Finally, the present study focuses, like most research, on written skills (Falchikov & Goldfinch, 2000; Double et al., 2020). However, it is likely that the domain can influence both PF processing and PF use (Double et al., 2020; Li et al., 2016). Therefore, future studies could replicate the present research in different subject areas (e.g. mathematics, Alqassab, Strijbos & Ufer, 2017) to determine whether there are differences in processing and use of PF depending on the subject area.

Notwithstanding the limitations as outlined above, the present article is the first to examine how discrepant PF is processed. Moreover, next to the previously used measure for revision performance, we included a second measure, which mimics higher education practices for assessing student work more closely. This allowed to discern that the operationalisation of revision performance impacts the substantive conclusions reached. Therefore, future studies in the field should use a better measure of performance in line with our operationalisation of it. In addition, the present study responded to the call for more "in vivo" research on feedback use (Jonsson, 2013; van der Kleij & Lipnevich, 2020): participants revised on-screen, which allowed us to track the gaze behaviour and analyse the (order of the) revision actions during the PF use subphase.

All in all, our results show that participants differed in how they processed discrepant PF: some noticed it but failed to report it afterwards, while others actively tried to make sense of it by transitioning between the discrepant PF elements. In addition, important differences were noted in how PF was acted upon: participants with a high-quality revision spent more time addressing higher order aspects, addressed one or two lower order aspects throughout the entire essay before tackling another aspect and conducted a proofreading stage at the end, in which they went beyond the feedback of their peers. These findings could be helpful when designing tools or training to foster students' processing and use of (discrepant) PF.

References

Aben, J., Timmermans, A., Dingyloudi, F., & Strijbos, J-W. (2021). *The effects of perceived language skills on peer feedback and peer grading in secondary education*. Paper presented at the Earli Conference 2021, Online.

Ahmadian, M., Yazdani, H., & Mehri, E. (2019). The effectiveness of learners' preferred and unpreferred written corrective feedback: a think-Aloud study. *The Journal of AsiaTEFL, 16*(2), 448-467. doi: 10.18823/asiatefl.2019.16.2.1.448

Alqassab, M., Strijbos, J.-W., & Ufer, S. (2017). Training peer-feedback skills on geometric construction tasks: role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education, 33*(3), 11-30. doi: 10.1007/s10212-017-0342-0

Altonji, S. J., Baños, J. H., & Harada, C. N. (2019). Perceived benefits of a peer mentoring program for first-year medical students. *Teaching and Learning in Medicine, 31*(4), 445-452. doi:10.1080/10401334.2019.1574579

Aryadoust, V. (2016). Gender and Academic Major Bias in Peer Assessment of Oral Presentations *Language Assessment Quarterly, 13*(1), 1-24. doi:10.1080/15434303.2015.1133626

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher, 40*(1), 62-69. doi:10.1080/0142159X.2017.1391373

Atkinson, N., L., Massett, H., A., Mylks, C., McCormack, L. A., Kish-Doto, J., Hesse, B., W., Wang, M. Q. (2010). Assessing the impact of user-centered research on a clinical trial eHealth tool via counterbalanced research design, *Journal of the American Informatics association18*(24), 24-31. doi: 10.1136/jamia.2010.006122

Berndt, M., Strijbos, J.-W., & Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *European Journal of Psychology of Education, 33*(1), 31-49. doi:10.1007/s10212-017-0343-z

Blanchard, K., & Johnson, S. (2015). *The new one minute manager*. New York (NY): William Morrow.

Bolzer, M., Strijbos, J. W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer-feedback via eye-tracking: role of feedback-characteristics, fixation-durations and transitions. *Journal of Computer Assisted Learning, 31*(5), 422-434.doi:10.1111/jcal.12091

Bouwer, R., & Dirkx, K. (2021). *The eye-mind of processing feedback: Unravelling how students read and use feedback for revision*. Paper presented at the Earli Conference 2021, Online.

Bouwer, R., Lesterhuis, M., Bonne, P., & De Maeyer, S. (2018). Applying Criteria to Examples or Learning by Comparison: Effects on Students' Evaluative Judgment and Performance in Writing. *Frontiers in Education, 3*(86), 1-12. doi:10.3389/feduc.2018.00086

Braasch, J. L. G., Rouet, J.-F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition, 40*(3), 450-465. doi:10.3758/s13421-011-0160-6

Buckingham, L., & Aktuğ-Ekinci, D. (2017). Interpreting coded feedback on writing: Turkish EFL students' approaches to revision. *Journal of English for Academic Purposes, 26*, 1-16. doi:https://doi.org/10.1016/j.jeap.2017.01.001

Butler, D. L., & Winne, P. H., (1995). Feedback and self-regulated learning : a theorical synthesis. *Review of educational research, 65*(3), 245-281.

Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Lesterhuis, M., & Van den Bossche, P. (2018). How are learning strategies reflected in the eyes? Combining results from self-reports and eye-tracking. *British Journal of Educational Psychology, 88*(1), 118–137. https://doi.org/10.1111/bjep.12181

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. Journal of economic behavior & organization, 81 (1-8), 1-8. doi: 10.1016/j.jebo.2011.08.009

Cho, K., & Schunn, C. D. (2018). Finding an optimal balance between agreement and performance in an online reciprocal peer evaluation system. *Studies in Educational Evaluation, 56*, 94-101. doi:https://doi.org/10.1016/j.stueduc.2017.12.001

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An information system design theory for the comparative judgement of competences. In: *European Journal of Information Systems*, Vol. 27, no.2, p. 248-261. doi:10.1080/0960085x.2018.1445461.

Coertjens, L., Lesterhuis, M., De Winter, B., Y., Goossens, M., De Maeyer, S., Michels, N., R., M. (2021). *Improving self-reflection assessment practices: Comparative judgment as an alternative to rubrics*. In: *Teaching and* Learning in Medicine: an international journal, (2021). doi:10.1080/10401334.2021.1877709 (Accepted).

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd edition ed.). New Jersey, NJ: Lawrence Erlbauw Associates.

Cutumisu, M., Turgeon, K.-L., Saiyera, T., Chuong, S., González Esparza, L. M., MacDonald, R., & Kokhan, V. (2019). Eye Tracking the Feedback Assigned to Undergraduate Students in a Digital Assessment Game. *Frontiers in Psychology, 10*(1931). doi:10.3389/fpsyg.2019.01931

Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., & Weber, B. (2012). On the use of neurophysiological tools in IS research: Developing a research agenda for neuroIS. *Mis Quarterly, 36*(3), 679-702.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review, 32*(2), 481-509. doi:10.1007/s10648-019-09510-3

El Ebyary, K., & Windeatt, S. (2019). Eye tracking analysis of EAP Students' regions of interest in computer-based feedback on grammar, usage, mechanics, style and organization and development. *System, 83*, 36-49. doi:https://doi.org/10.1016/j.system.2019.03.007

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

Fong, C. J., Schallert, D. L., Williams, K. M., Williamson, Z. H., Lin, S., Kim, Y. W. & Chen, L.-H. (2021): Making feedback constructive: the interplay of undergraduates' motivation with perceptions of feedback specificity and friendliness. *Educational Psychology*, 1-19. doi: 10.1080/01443410.2021.1951671

Garino, A. (2020). Ready, willing and able: a model to explain successful use of feedback. *Advances in Health Sciences Education, 25*(2), 337-361. doi:10.1007/s10459-019-09924-2Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304-315. doi:https://doi.org/10.1016/j.learninstruc.2009.08.007

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304-315. doi:https://doi.org/10.1016/j.learninstruc.2009.08.007

Ginsburg, S., Vleuten, C. P., Eva, K. W., & Lingard, L. (2017). Cracking the code: residents' interpretations of written assessment comments. *Medical Education, 51*(4), 401-410. doi:doi:10.1111/medu.13158

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Hessel, A. K., & Schroeder, S. (2020). Interactions Between Lower- and Higher-Level Processing When Reading in a Second Language: An Eye-Tracking Study. *Discourse Processes, 57*(10), 940-964. doi:10.1080/0163853X.2020.1833673

Holmqvist, K., & Andersson, R. (2017). Eye tracking: *A comprehensive guide to methods, paradigms and measures*. Lund, Sweden : Lund Eye-Tracking Research Institute.

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2018). Peer feedback on academic writing: undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment & Evaluation in Higher Education, 43*(6), 955-968. doi:10.1080/02602938.2018.1424318

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing, 7*(3), 255-286.

Isohätälä, J., Näykki, P., & Järvelä, S. (2019). Cognitive and socio-emotional interaction in collaborative learning: exploring fluctuations in students' participation. *Scandinavian Journal of Educational Research*, 831-851. doi: 10.1080/00313831.2019.162331

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151-177. doi:10.1007/s10763-013-9497-6

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63-76. doi:10.1177/1469787412467125

Kaakinen, J., K., & Hyönä, J. (2007). Strategy use in the reading span test: An analysis of eye movements and reported encoding strategies. *Memory, 15*(6), 634-646. doi: 10.1080/09658210701457096

Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restor Den Endod., 42*(2), 152–155. doi : 10.5395/rde.2017.42.2.152

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155-163. doi:10.1016/j.jcm.2016.02.012

Kyaruzi, F., Strijbos, J.-W., Ufer, S., & Brown, G. T. L. (2019): Students' formative assessment perceptions, feedback use and mathematics performance in secondary schools in Tanzania, *Assessment in Education: Principles, Policy & Practice, 26*(3), 278-302. doi: 10.1080/0969594X.2019.1593103

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., et al. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. Assessment & Evaluation in Higher Education, 41 (2), 245-264. DOI 10.1080/02602938.2014.999746

Lipnevich, A., Murano, D., Krannich, M., & Goetz, T. (2021). Should I grade or should I comment: Links among feedback, emotions, and performance. *Learning and Individual Differences, 89*, 1-19. doi: 10.1016/j.lindif.2021.102020

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Máñez, I., Vidal-Abarca, E., Kendeou, P., & Martínez, T. (2019). How do students process complex formative feedback in question-answering tasks? A think-aloud study. *Metacognition and Learning, 14*(1), 65-87. doi:10.1007/s11409-019-09192-w

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282. doi: 10.11613/BM.2012.031

Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia

learning. *Computers in Human Behavior, 96*, 223-234. doi:https://doi.org/10.1016/j.chb.2018.06.028

Penttinen, M., Anto, E., & Mikkilä-Erdmann, M. (2013). Conceptual change, text comprehension and eye movements during reading. *Research in Science Education, 43*(4), 1407-1434. doi: https://doi.org/ 10.1007/s11165-012-9313-2

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594X.2012.665354

Prilop, C. N., Weber, K. E., Prins, F. J., & Kleinknecht, M. (2021). Connecting feedback to self-efficacy: Receiving and providing peer feedback in teacher education. *Studies in Educational Evaluation*, 70, 1-12. doi: 10.1016/j.stueduc.2021.101062

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading, 10*(3), 241-255. doi:10.1207/s1532799xssr1003_3

Richardson, J. T. E. (2013). Research issues in evaluating learning pattern development in higher education. Studies in Educational Evaluation, 39(1), 66-70. doi: https://doi.org/10.1016/j.stueduc.2012.11.003

Rinck, M., Gámez, E., Díaz, J. M., & De Vega, M. (2003). Processing of temporal information: Evidence from eye movements. *Memory & Cognition, 31*(1), 77-86. doi:10.3758/BF03196084

Schoor, C., Rouet, J.-F., & Britt, M. A. (2021). *Context and consistency effects when reading multiple documents.* Paper presented at the Earli Conference 2021, Online.

Schüler, A. (2017). Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration. *Learning and Instruction, 49*, 218-231. doi:https://doi.org/10.1016/j.learninstruc.2017.03.001

Schüler, A. (2019). The integration of information in a digital, multi-modal learning environment. *Learning and Instruction, 59*, 76-87. doi:https://doi.org/10.1016/j.learninstruc.2017.12.005

Stadtler, M., Scharrer, L., & Bromme, R. (2020). How Relevance Affects Understanding of Conflicts Between Multiple Documents: An Eye-Tracking Study. *Reading Research Quarterly, 55*(4), 625-641. doi:https://doi.org/10.1002/rrq.282

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing With Uncertainty: Readers' Memory for and Use of Conflicting Information From Science Texts as Function of Presentation Format and Source Expertise. *Cognition and Instruction, 31*(2), 130-150. doi:10.1080/07370008.2013.769996

Strijbos, J.-W., & Wichmann, A. (2018). Promoting learning by leveraging the collaborative nature of formative peer assessment with instructional scaffolds. *European Journal of Psychology of Education, 33*(1), 1-9. doi:10.1007/s10212-017-0353-x

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). *Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus*. In: *Assessment in Education: Principles, Policy and Practice*, Vol. 26, no. 1, p. 59-74 (2019). doi:10.1080/0969594X.2016.1253542

van der Kleij, F. M. (2020). Evaluation of the 'Feedback Engagement Enhancement Tool' to examine and enhance students' engagement with feedback on their writing. *Studies in Educational Evaluation, 66*, 100907. doi:https://doi.org/10.1016/j.stueduc.2020.100907

van der Kleij, F. M., Lipnevich, A. A. (2020). Student perceptions of assessment feedback: a critical scoping review and call for research. *Educationnal Assessment, Evaluation and Accountability, 33*(2), 1-29. doi: 10.1007/s11092-020-09331-x

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance: combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology, 19*(2), 205-221.doi:10.1002/acp.1112

van Moort, M. L., Koornneef, A., & van den Broek, P. W. (2021). Differentiating Text-Based and Knowledge-Based Validation Processes during Reading: Evidence from Eye Movements. *Discourse Processes, 58*(1), 22-41. doi:10.1080/0163853X.2020.1727683

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. Learning and Individual Differences, 29, 123-130. doi: https://doi.org/10.1016/j.lindif.2013.01.003

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541-562. doi:10.1080/0969594X.2019.1602027Wichmann, A., Funk, A., & Rummel, N. (2018). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *European Journal of Psychology of Education, 33*(1), 165-184. doi:10.1007/s10212-017-0348-7

Winstone, N. E., Hepper, E. G., Nash, R. A. (2021). Individual differences in self-reported use of assessment feedback: the mediating role of feedback beliefs. *Educational Psychology, 41*(7), 844-862. doi: 10.1080/01443410.2019.1693510

Winstone, N.E., Mathlin, G., & Nash, R. A. (2019). Building Feedback Literacy: Students' Perceptions of the Developing Engagement With Feedback Toolkit. *Frontiers in Education, 4*(39), 1-11. doi: 10.3389/feduc.2019.00039

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. *Educational Psychologist, 52*(1), 17-37. doi:10.1080/00461520.2016.1207538

Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. *Studies in Higher Education, 42*(11), 2026-2041. doi:10.1080/03075079.2015.1130032

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology, 10*(3087). doi:10.3389/fpsyg.2019.03087

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179-200. doi:https://doi.org/10.1016/j.jslw.2006.09.004

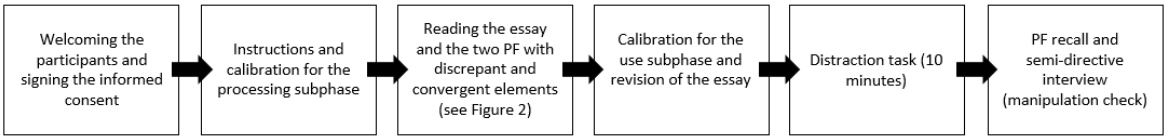

Figure 1. Overview of the experiment

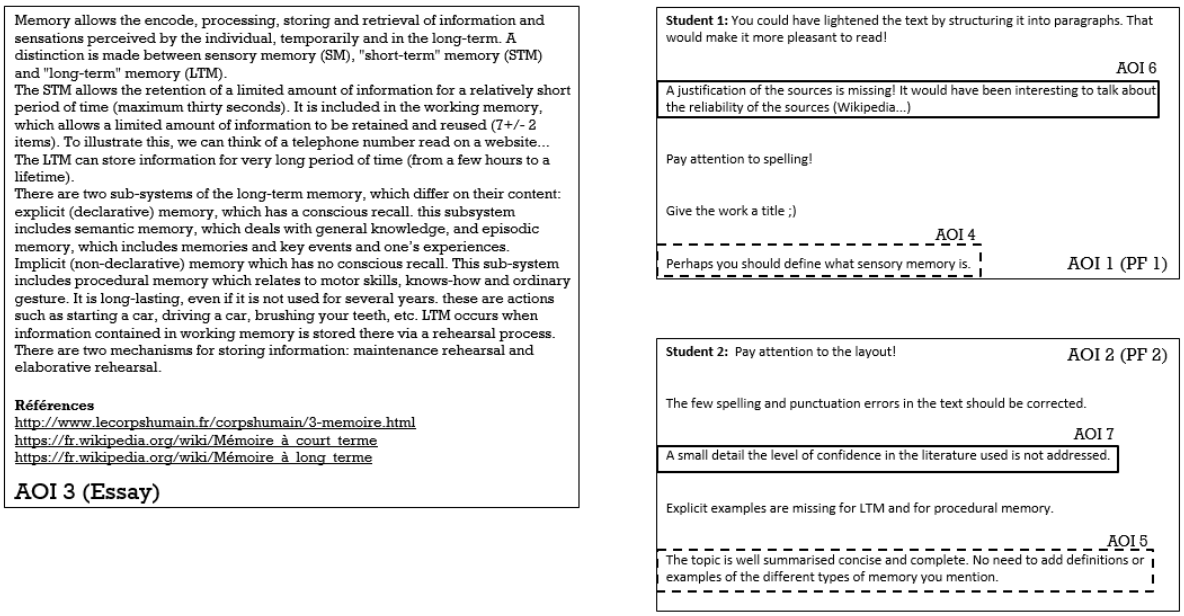

Figure 2. Essay and PFB, with the AOIs

Note: In the original experiment the essay and the peer feedback were presented in French

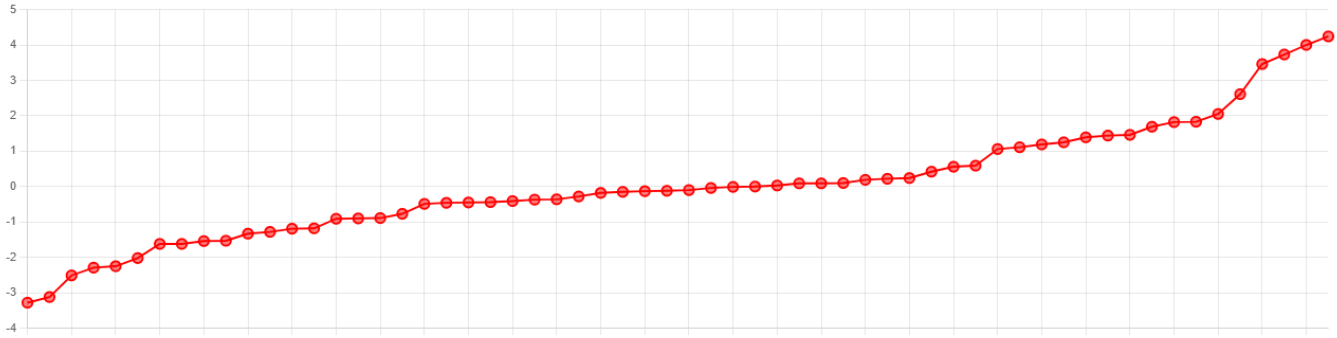

Figure 3: rank order of the revision performance produced by the Comproved software.

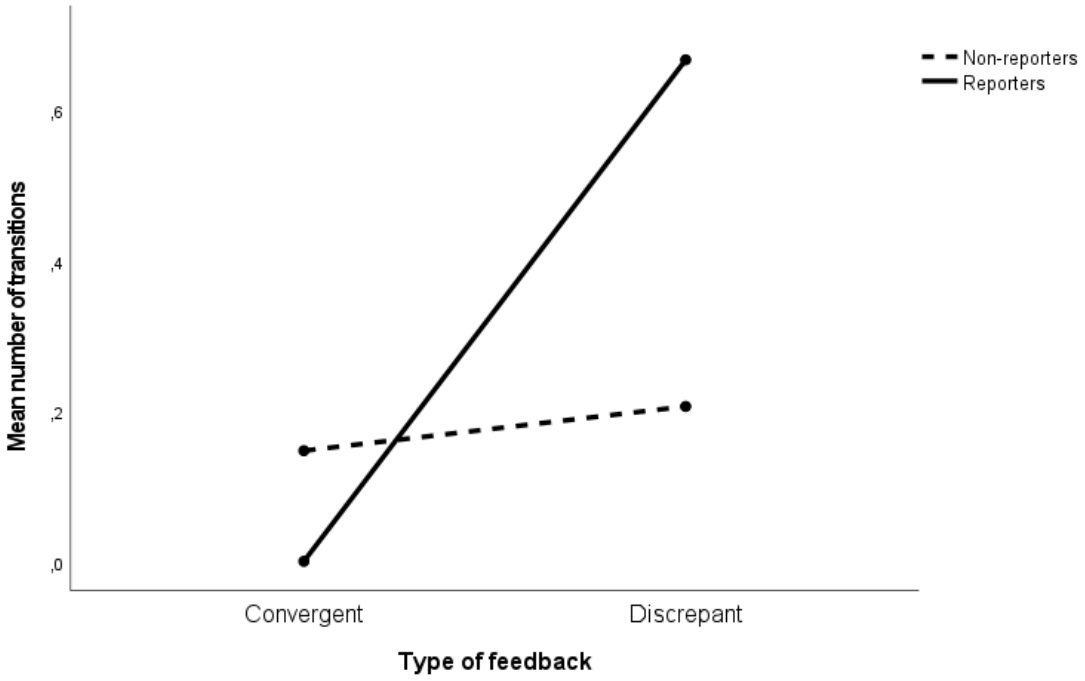

Figure 4. Mean number of transitions between peer feedback elements as a function of discrepancy (Convergent *vs.* Discrepant) and Reporting Group (Reporters *vs.* Non-reporters). Reporters significantly differed from Non-reporters on the number of transitions for discrepant peer feedback elements.

Table 1.

Coding scheme for peer feedback use

	(Not) in line with peer feedback elements	**Main code**	**Subcode**	**Lower or higher order aspect**
Reading	/	Reading	Reading the text Reading the peer feedback	
Revising	In line with peer feedback elements	Correction of spelling	Spelling Punctuation	Lower order aspect
		Lay-out	Creating paragraphs	Lower order aspect
		Adding a title	/	Lower order aspect
		Reliability of the references	/	Higher order aspect
		Definition and examples	Sensory memory Examples of long-term memory Examples of procedural memory	Higher order aspect
	Not in line with peer feedback elements	Lay-out	Connecting words and pronouns Revising sentences Adding subheadings Highlighting keywords or putting them in bold	Lower order aspect

Table 2.

Overview of the different actions and their sequence for one of the 10 participants with a high-quality and one of the 10 participants with a low-quality revision

Participant with a high-quality revision	**Participant with a low-quality revision**
Reading the PF	Reading the PF
Making spelling and punctuation corrections for the entire text	Creating paragraph 1
Adding a title	Making spelling and punctuation corrections for paragraph 1
Creating paragraphs throughout the entire text	Making spelling and punctuation corrections for paragraph 2
Adding the definition of sensory memory	Creating paragraph 2
Changing the structure of sentences	Making spelling and punctuation corrections for paragraph 3
Changing the title anew	Creating paragraph 3
Adding examples of long-term memory	
Adding a note on the reliability of the references	
Proofreading stage	

Appendix A

Guidelines for students and assessment criteria

Guidelines

As we discussed during the course "General psychology: processes & theories", several types of memory can be mobilised by human beings. Referring to the concepts seen in class, you are asked to write a text of about 20 lines that summarises the knowledge you gained on this topic. This text will then be read by two of your colleagues who will judge it according to different criteria (see below).

Assessment criteria

Criterion 1 : This text gives clear and comprehensible information
Criterion 2 : this text gives detailed information
Criterion 3 : this text makes at least 3 ties with contents covered during the course "General Psychology: processes & theories"
Criterion 4 : the author of this text follows the instructions
Criterion 5 : the author of this text adequately summarized the information
Criterion 6 : the author of this text cites the bibliographical sources used
Criterion 7 : I found the text interesting to read
Criterion 8 : The length of the text was good
Criterion 9 : I was impressed by this text

Appendix B

Questions asked during the semi-directive interview

- Can you tell me what you thought of the task?
- Can you tell me how you went about reading the work and the feedback and revising the work?
- Did you find the task difficult or easy?
- Did you notice any contradictions between the feedbacks of the peers? If so, can you tell me how you dealt with them?
- To what extent do you feel the revised work is of better quality?
- Is there anything you would like to add?