Do Students Generate Better Self-Feedback by Comparing their Work Against Assessment Criteria or Exemplars?

Maxine V. Swingler, School of Psychology and Neuroscience, University of Glasgow*

David Nicol, Adam Smith Business School, University of Glasgow

Lorna Morrow, School of Psychology and Neuroscience, University of Glasgow

Author Note

Maxine V. Swingler https://orcid.org/0000-0002-0108-0212

We have no known conflict of interest to disclose

Correspondence concerning this article should be addressed to: Maxine Swingler, School of Psychology and Neuroscience, University of Glasgow, G12 8QB. Email: Maxine.swingler@glasgow.ac.uk

Abstract

Having students write their own feedback comments can positively impact their performance. To support these self-feedback productions, teachers usually provide students with information to compare their work against. Recent studies show that both assessment criteria, often in the form of a rubric, and exemplars promote grade improvements in draft-redraft situations. Yet little is known about the differential effects of these comparators on learning itself. The current study investigated this in a peer review setting. Students produced a written research report then one group compared it against information in the assessment criteria and another against information in exemplars. Both groups wrote self-feedback comments based on the comparisons they made. Results show that students produced more high-level process comments when comparing against exemplars while they produced more task-related feedback comments when comparing against assessment criteria. Students' final grade performance was also better after exemplar comparisons. The implications of using different comparators to promote different kinds of self-feedback are discussed, as is the value of using instructional prompts to target different learning outcomes. By extending the range of comparators beyond the limited set currently in use and through appropriate prompting, teachers can help shift feedback agency to students and reduce their own workload.

Keywords: internal feedback, assessment rubric, exemplars, peer review, instructional prompts, student feedback agency, self-feedback, self-regulation, feedback comparisons.

Introduction

Over the last 10-15 years research on feedback has evolved. Initially focused on what teachers do, on the quality and timeliness of the feedback comments they provide, attention has now shifted to what students do, to how they engage with these comments (Boud and Molloy, 2013; Van der Kleij et al., ,2019). Engagement encompasses how students seek out, elicit, process, and interpret comments from others as well as how they use them to make performance improvements (Carless & Boud, 2018; de Kleijn, 2023; Molloy et al., 2020; Winstone et al, 2017). This shift in student feedback agency is framed within a dialogic conception of feedback, where feedback is viewed as a communicative exchange. While the importance of feedback dialogue is incontrovertible, new research suggests that conceiving feedback through this lens alone is limiting, as it ignores the feedback processes that students naturally engage in, for example, when consulting a textbook or an online resource (Jensen et al., 2022; Nicol, 2021; Smith et al., 2023).

Researchers investigating how to capitalise on such natural feedback processes have shown that the information teachers provide to students about their performance need not always comprise comments (Burnell et al, 2023; Lipnevich et al, 2014; Nicol and Kushwah, 2023; Tomazin, et al, 2023). Teachers can also provide students with information in other resources, for example, in exemplars, assessment criteria and rubrics and ask them to consider their work through the lens of this information. When students are asked to do this, they compare their work against the information they derive from the resource and activate an inner feedback process, which leads to changes in their knowledge and performance (Nicol, 2021; Panadero et al, 2019; Vreug et al., 2023). Nicol (2021) provides a feedback model based on these ideas, and advice to help teachers design comparison activities to help build students' natural inner feedback capability (Nicol, 2022). A key principle in the model is that students must make mindful comparisons and make their inner feedback processes explicit, for example, by writing self-feedback comments.

One important implication deriving from the inner feedback model is that when teachers vary the information they assign for comparison, students will activate different self-feedback knowledge and produce different kinds of self-feedback comments (Nicol, 2021). For example, one would expect that when students compare their work against a rubric, they would produce different feedback comments than when comparing against exemplars, as each embodies different information. "Rubrics provide information on the goals of the assignment while exemplars provide models of what a desired final state would be" (Burnell et al., 2023, p2). Stated differently, rubrics provide tacit information about the quality of the expected work whereas exemplars provide concrete representations of what quality looks like (Rust et al, 2003; Sadler, 2010).

Where assessment rubrics and exemplars have been used as feedback comparators in the same study, results show that rubrics are more effective than exemplars (Lipnevich et al, 2014: Lipnevich et al, 2023). However, in these studies the measure of effectiveness was grade improvements in a draft-comparison-redraft situation. Hence, these studies do not provide direct evidence about the internal feedback processes (i.e., changes in knowledge) that students activate from each comparison, rubric or exemplars. This is a critical omission, as many researchers argue that performance (i.e., achieving good grades) and learning are not synonymous (Wiliam, 2018; Burnell et al, 2023) and call for outcome measures that are more illuminating of learning than just grades (Winstone and Nash, 2023). Also, where exemplars have been used as comparators, with students making visible in writing their own feedback productions, they produced self-feedback comments on higher-order aspects of their work rather than on low level aspects (e.g. Nicol and

McCallum, 2022: Nicol and Kushwah, 2023). Examples of the former include feedback on the structure of the argument, on the evidence argument relationship and on how readers might interpret the students' work while examples of the latter include correcting errors in subject content or in spelling and grammar (Hattie & Timperley, 2007; Nicol & Selvaretnam, 2021). This suggests that students are engaging in deep rather than surface learning when making exemplar comparisons. However, we could find no published studies where students produce self-feedback comments from rubric comparisons.

In this study we directly examine what self-feedback students produce when they compare their work against information in assessment criteria versus information in exemplars, as well as how each comparator influences their performance, their final assessment grade. Assessment criteria were used rather than an assessment rubric to focus students on the writing of self-feedback comments rather than on self-assessing their work, which is one of the main reasons for using a rubric (Andrade, 2019; Taylor et al., 2024). By varying the comparison information beyond a single focus on teacher comments, educators can broaden the scope of the feedback students produce while also building students' natural feedback agency. Further, as providing comparators as feedback information is likely to be less time consuming than writing comments, this research also has implications for teacher workload.

Assessment Criteria, Rubrics, Exemplars, and Self-Feedback

Exemplars, assessment criteria and rubrics have been widely used in educational research to support students' learning. A rubric differs from assessment criteria in that it lists the criteria for different levels of quality (Panadero and Jonsson, 2013; Brookhart, 2018). In most studies, these resources have been used *before* students produce work to help clarify the task or assessment requirements rather than after they produce work as a tool to generate self-feedback (Panadero & Lipnevich, 2021; To et al, 2021). Hence, with one exception, Bouwer et al (2018), this 'before' research is not reviewed here.

Assessment Rubrics and Exemplars: Performance Measures. Studies where rubrics or exemplars have been used as feedback information show that these result in differential improvements in grades in a draft-redraft situation. These studies thus provide indirect evidence for the claim that different comparators result in differential changes in students' knowledge. Lipnevich et al (2014), for example, had students write a draft research report then assigned different groups to one of three comparison conditions - a detailed rubric, three exemplars of performance and a combination of rubric and exemplars. After making comparisons students updated their drafts. All three groups showed significant improvement from draft to redraft with the

rubric group making most improvement. However, one confounding factor was that the rubric was also used by the instructor to grade the students' performance. Further, the poorer performance by students in the combined rubric and exemplar condition versus the rubric-alone condition was puzzling, as one would think that more comparators would result in more learning. Lipnevich et al's (2014) interpretation is that students had to complete two tasks, within the same time scale as those who did one task, so the cognitive load was greater.

In a follow-up study, Lipenvich et al (2023) studied ninth and tenth grade students in similar situation with rubrics, exemplars, a combination of both, and a control condition and the findings were the same. Yet when the instructor explained to students how to use exemplars for comparison the differences between the rubric and exemplars condition was less. This suggests that the instructions teachers give to students might mediate what they learn from making feedback comparisons (Narciss, et al., 2022; Nicol, 2022).

Exemplars versus Teacher Comments. Exemplars have also been compared against teacher comments. Tomazin et al (2023), for example, found no difference in performance in a draft-redraft situation where one group of students received annotated exemplars as comparators after writing an essay and another received individualised teacher feedback comments (see also, Price et al., 2017). This result shows that comments are not always the better comparator in terms of grade improvements, a finding that has significant implications for teacher workload. There are however compelling reasons to think that teacher comments and exemplars would lead to differential changes in students' knowledge and that the performance measures in this study are not picking this up. Comparisons against comments involve students in comparing their work against a judgement made by others of that work whereas comparisons against exemplars call on students' judgements, relatively unconstrained by the judgement of others (Carless and Boud, 2018; Nicol and Kushwah, 2023). Also, comments give information about what needs improving while exemplars provide concrete illustrations of different kinds of work.

Performance and Learning. Burnell et al (2023) have recently critiqued studies on assessment rubrics and exemplars as feedback comparators, arguing that improvements in short-term performance in a draft-redraft situation tell us little about student learning. These researchers define learning as 'what learners take with them that can be used on a subsequent task' (p1). In their study, students learning English as a second language wrote a 350-word essay then made comparisons against different types of information and then revised their submitted essays. One group compared their essays against a rubric, another against an exemplar

on a different topic, a third group (self-assessment group) rated their performance against the assessment criteria and wrote about the strengths and weaknesses in their essays. In a control condition, students just re-wrote their essays. To test long-term learning, students wrote a new essay two weeks later. Results showed that in terms of short-term performance the exemplars were more effective than the rubric but in the long term the rubric was more effective than the exemplars.

While this study is important, arguably it still confounds learning and performance, only now learning is defined as long-term performance. Further, that the rubric was less effective than the exemplars in the short-term, draft-redraft situation contradicts the findings of Lipnevich and her colleagues (Lipnevich et al 2014: Lipnevich et al, 2023). An alternative way of investigating learning from different kinds of feedback information is to surface the internal processes of learning in the short term, rather than relying on performance in the long term. One way to do this is to have students write their own self-feedback comments based on the comparisons they make (Nicol, 2021; Nicol, 2022).

Nicol and McCallum (2021), for example, report that when students compared their essays directly with other students' essays, and wrote self-feedback comments, they generated additional self-feedback beyond that stated in the assessment criteria (e.g., motivational, reader perspectives). In turn, Nicol & Selvaretnam's (2022) found that when students compared their individual work against the same work subsequently produced by their group (in a two-stage exam) they generated high levels of *process* and *self-regulatory* feedback (e.g., on how they might better plan future work). Similarly, Nicol and Kushwah (2023) report that when students made comparisons of their dissertation literature reviews against published research literature reviews on a different topic, they generated self-feedback on the quality of their own writing (e.g. the evidence-argument relationship) rather than on the subject content. Each of these studies suggest that students are generating feedback at a holistic level on relational aspects of their work rather than at the task level on the subject content (Hattie & Timperley, 2007, see also, Sadler, 2010). Unfortunately, no studies have so far been conducted where students made explicit the self-feedback that they generated from comparisons against assessment criteria or rubrics.

Writing Comments for Peers. There is however one relevant study by Bouwer et al (2018) in which students wrote feedback comments on their peers' work (rather than their own work) either using assessment criteria or exemplars as comparators. In this study, before producing their own essay, students evaluated essays written by peers the year before. One group compared the peer essays against each other (comparative

judgement condition) and wrote feedback on the essays they compared. The other group compared the peer essays against assessment criteria and wrote feedback. The exemplar comparison condition (i.e. peer essay comparison) resulted in students producing feedback comments on higher-order aspects of the text whereas the assessment criteria comparison condition led to more specific comments on grammar and vocabulary. Higher order here was defined as feedback that calls for changes to the overall meaning or purpose of the work, also identified as global feedback, as opposed to local or surface level feedback (Underwood & Tregidgo, 2006). These differences did not translate into significant differences in students' performance in their own essay writing, although the researchers suggest that this might be due to the small number of participants. Nonetheless, these findings resonate with those of Nicol and others where exemplar comparisons also resulted in students writing self-feedback comments at a high meaning level (e.g. Nicol and McCallum, 2021).

In sum, studies to date suggest that assessment criteria (usually in the form of rubrics) and exemplars lead students to produce different kinds of self-feedback comments as these resources contain different information. Yet the findings are incomplete. On the one hand, when students compare their draft work against the information provided by exemplars and rubrics in a draft-redraft situation, the result is differences in grade improvements with rubrics usually superior. However, grade improvement studies say little about the learning process, about what feedback processes and knowledge students are activating internally. In contrast, research where students compare their work against information in exemplars and make their inner feedback processes explicit as self-feedback comments seems to suggest that exemplars lead students to generate higher order feedback at a relational and deep learning level. Yet there are no comparable studies where students have produced self-feedback from assessment criteria or rubrics.

The Current Study

The purpose of the current study was to investigate what students learn from comparing their work against assessment criteria versus exemplars. Unlike Burnell (2023), we define learning as what feedback students self-produce while making comparisons rather than in terms of their subsequent performance in a new task. In other words, making their feedback processes visible as self-feedback comments provides direct evidence of student's learning as it is happening. Hence the first research question:

1) Do students produce different self-feedback comments when they compare their work against assessment criteria versus exemplars?

To examine differences in students' self-feedback comments, we used Hattie and Timperley's (2007) teacher-feedback framework to categorise students' self-feedback comments. That framework distinguishes between task level feedback and process level feedback. Task feedback comprises the comments that teachers provide on students' performance on the task for example, on the strengths and weaknesses in the students' work and on what could be improved. Process level feedback refers to feedback comments on the processes used to complete the task such as feedback on how students conceive their work holistically, how they inter-relate different elements of their work and on the thinking processes they engage in to complete the task.

Hattie and Timperley (2007) found that process feedback from a teacher is far more effective than task feedback in promoting deep learning and learning transfer to new tasks (see also, Wisniewksi et al., 2020). Yet it has also been found that most teacher feedback comprises task level comments (Arts et al, 2016; Derham et al 2021; Dirkx et al, 2021). If Hattie & Timperley's (2007) distinction between task and process level feedback has a parallel in terms of students' own self-feedback productions, then this would be important to understand, as it would enable teachers to design feedback comparison activities that are more likely to lead to deep learning by students and learning transfer.

This study was also designed to provide data on students' performance as grades, as a result of making feedback comparisons. However, we did not compare grades between draft and redraft as in Lipnevich et al (2023) nor did we follow the pattern of Burnell et al (2023) of investigating performance in a subsequent task. Instead, we examined students' performance on their *final submitted work*. There were two measures: (i) students' grade on their final work (a research report) which was submitted two weeks after making comparisons against the criteria or exemplars and writing self-feedback; (ii) the relationship between students' grade and the type of self-feedback comments they produced. Performance was thus measured in light of the final outcome *and* in relation to the self-feedback process. Hence the second research question was in two parts:

2)

a) How do different types of comparators, assessment criteria or exemplars, impact students' final assessment grade?
b) How does the type of self-feedback comments generated by students relate to their final assessment grade?

These research questions were investigated in the context of a peer review activity. Also, instructional prompts were used to focus students on relevant information in the comparators and to guide their written self-feedback productions (Nicol and Kushwah 2023)

Method

Participants

Participants were 113 undergraduate (UG) and postgraduate taught (PGT) psychology students enrolled in a qualitative methods course at a UK University. Fifty-seven of these students participated in the course in one academic year (cohort 1), and 56 participated in the subsequent academic year (cohort 2). Participant demographic characteristics are detailed in Table 1. Participants were screened to ensure they had participated in all elements of the self-feedback process and that their final grades for the course represented the range of grades achieved by all students in their year group. The sample size met minimum requirements for an independent t-test at the .05 probability level, with a power level of .8 and anticipated effect size of d = 0.5, which was 102 (Soper, 2023).

Table 1.

Participant Characteristics in Cohort 1 and Cohort 2.

	Cohort 1: comparison with assessment criteria		Cohort 2: comparison with exemplars	
	UG	PGT	UG	PGT
Sample for coding comments	31	26	31	25
% Female	84	77	97	88
% Male	16	23	3	12
% UK/EU students	87	58	97	54
% International students	13	42	3	46

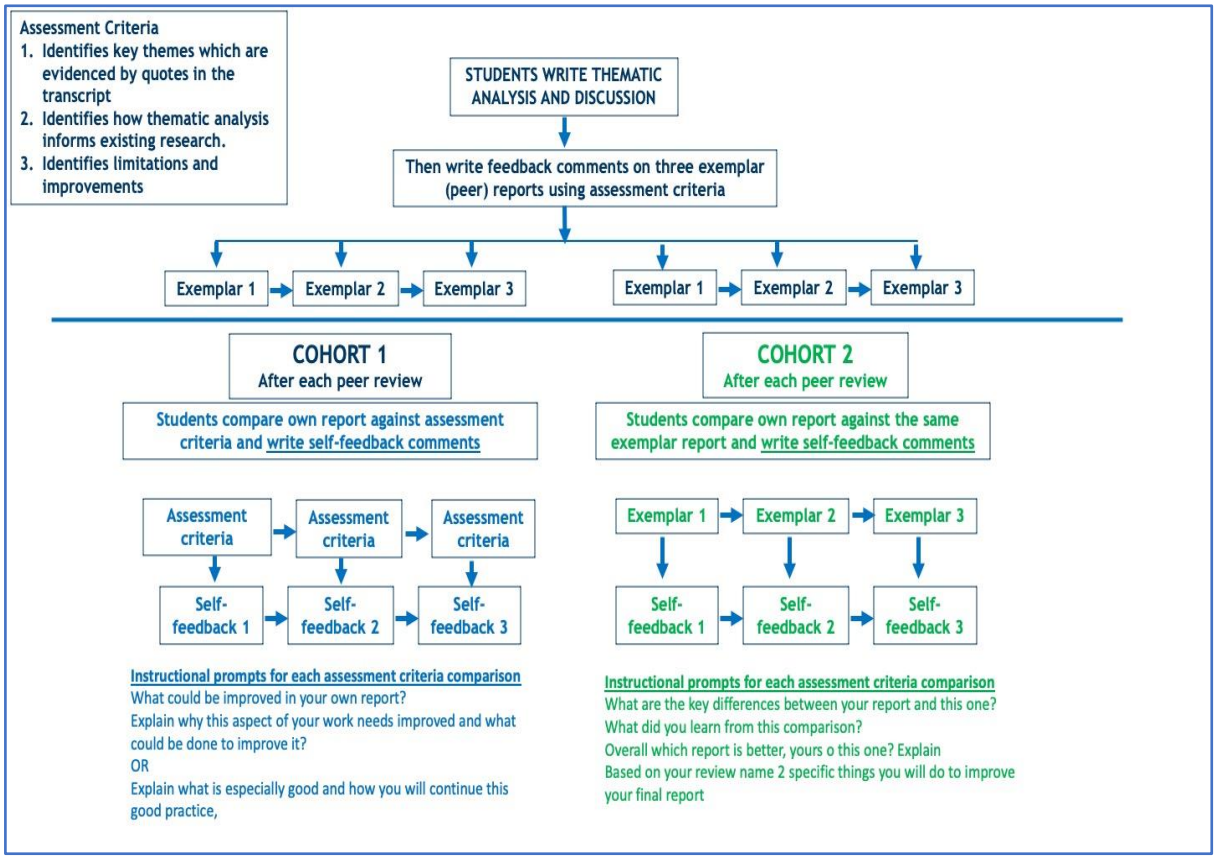
Participant recruitment, procedures, and data collection methods followed the guidelines outlined in the British Psychological Society code of Human Research Ethics (BPS, 2016). The research was approved by the University Ethics committee (Ref: 300190080).

Procedure

Students carried out a qualitative psychology research project over 11 weeks, including data collection, analysis, and the writing up of a research report. Students in both cohort 1 and cohort 2 produced a short formative report of their analysis and discussion which was used as the basis of a subsequent peer- and self-feedback task. The focus of this research is the self-feedback task. The peer feedback task was identical for both cohorts. This parallels the design used by Nicol and McCallum (2022). An outline of the sequence of the peer and self-feedback tasks is presented in Figure 1.

Figure 1.

Outline of the procedure for the exemplar and criteria comparisons and the instructional prompts used for each comparison group, cohort 1 (assessment criteria) and cohort 2 (exemplars).



Students in both cohorts anonymously peer reviewed three reports and wrote feedback comments about them (peer feedback) using the assessment criteria to frame their comments. The assessment criteria are indicated in Figure 1 (numbered 1, 2 and 3). Aropa software (Purchase and Hamer, 2018) was used to redistribute peer reports and provide a space for students to write their peer and self-feedback comments. Two of the peer reports were produced by classmates and one was an exemplar produced by the teacher. For the purpose of this study, *exemplars* refer to both to peer works (exemplars 1 & 2 in Figure 1) *and* the work produced by the teacher (exemplar 3 in Figure 1). All exemplars were on a similar research topic and based on analysis of data that students had collected themselves.

After each peer feedback review, students made comparisons of their own report against the relevant comparator and wrote self-feedback comments. One cohort (cohort 1) compared their work against each of the three assessment criteria and wrote self-feedback comments guided by the instructional prompts shown in Figure 1 (bottom left). The other cohort (cohort 2) compared their work against the exemplar (they had just

reviewed) and wrote feedback comments, guided by the instructional prompts shown in Figure 1 (bottom right). In effect, all students had three opportunities to write self-feedback comments. After writing their own self-feedback both cohort 1 and cohort 2 incorporated what the learned from this into their final report which was submitted 2 weeks later for summative assessment.

Measures

Coding of Students' Written Self-Feedback Commentaries. The unit of analysis used in coding of self-feedback was a phrase or sentence in the self-feedback commentary with a distinct meaning (Maguire & Delahunt, 2017). This coding was both deductive and inductive. In order to directly compare the self-feedback comments between cohorts 1 and 2, we adapted Hattie & Timperleys' (2007) categorisation of task and process-level feedback, an approach utilized in recent studies of feedback commentaries (Derham et al., 2021; Dirkx et al., 2019; Nicol and Selvaratnam, 2021).

Table 2

Coding categories for self-feedback comments.

Category of internal feedback	Definition and scope
Task	Comments on performance on the given task. Comments on strengths and weaknesses in the content, errors, inaccuracies, or missing elements in their work
Process	Comments on the processes and strategies used to perform the task. Comments on the inter-relationships across components of their work (e.g., the research questions and analysis, themes and supporting quotes, results and discussion), structure and flow of the argument, use of evidence to support the argument, and how readers might interpret their work.
No elaboration	Provides very brief task or process comment that does not explain anything in detail or give an example.
Elaboration	Provides detailed explanation (task or process) of aspects of their work and/or a specific example.

After an initial phase of coding, the task and process categories were deemed suitable in capturing the self-review comments. However, due to variability in the level of detail of student comments, the coding scheme was refined so that comments were also categorised according to their degree of elaboration (elaboration or little/no elaboration) leading to 4 categories of self-review comments: i) Task – non-elaboration; ii) Task-elaboration; iii) Process-non elaboration; iv) Process-elaboration. Table 2 provides definitions and examples of the coding scheme and Table 3 provides examples of students' comments in each category.

Table 3.

Examples of students' self-review comments, in each category, selected from coded comments from the assessment criteria group and the exemplar group.

Category of internal feedback	Examples- little/no elaboration	Examples- elaboration
Task	*"..find more supporting quotes"* *"My analysis had 5 quotes, the reviewed analysis had 3"* *"Maybe break the research question down if necessary."*	*"…love the use of direct emotions being used as potential themes. could be potential themes to discuss in my own submission as boredom, relaxation etc are growing themes in my research report too."* *"I mentioned the limitation that our focus group generated very little material to directly engage with our initial research question. I could say more about this and then actually only include the refined research question."* *The structure of themes was very good and I could improve mine by cutting down*

		some of my quotes and just including page numbers.
Process	*"I have used this to re-read my own work, and make sure it is understandable to others, ensuring it follows a clear structure, like this."* *"I learnt the importance of clearly marking out what your themes are, as otherwise the reader may interpret what is written differently".* "..having really good supporting literature for discussion really helps to contextualise the research and how it can contribute to future policies and research."	*"I have learnt from this peer's TA (Thematic Analysis) that the structure of the write up can have a huge impact in getting across the message effectively to the reader. The discussion should focus primarily on relating the results to other research, whereas the analysis should include more narrative and explanation of results."* *Show how quotes support discussion, through for instance words that convey barriers for example to help relate back to the research question. note that there can be positive and negative aspects of the same phenomenon.* *First would be the interpretation itself. I would like to go through the transcription again taking a more active role to see if I could delve something more in-depth and implicit. By combining quotes from different participants together also enhance the validity of the analysis. .*

An informal pilot coding involved two of the authors independently coding a sample of 66 student comments from cohort 1 and cohort 2 using a draft version of the coding scheme. A Cohens Kappa coefficient of .76 was obtained which is appropriate for exploratory studies (Lombard et al., 2010).

Clarifications were discussed for any discrepancies and the coding scheme was refined (see Table 2). The first author then proceeded to code comments from all participants (total comments = 1423). For reliability checking, 80 comments were randomly selected and independently coded by the first and second authors, and a Cohens' Kappa of k = .78 was obtained.

Performance Grades. The final qualitative reports for the participant sample were graded on a scale from 0 to 22 according to the university code of assessment. Experienced teaching staff (including one member of the research team) graded the assessments, and moderation procedures were in accordance with university guidelines. Any moderated reports with a difference of more than 2 grade points between markers were discussed and the discrepancy resolved. Participants' grades for their final report and their Grade Point Average (GPA) on a 22-point scale, representing performance on all assessments in the academic year, were included in the analysis (see Table 4).

Results

Preliminary Analysis

Assumptions of normality in the distribution of self-feedback comments in each of the 4 categories (described in Table 2) were not met, thus, inferential tests of Research Question 1 were non-parametric. Effect sizes for post-hoc comparisons were calculated by converting the test statistic into a z score, then calculating an effect size estimate, r, from the z-score. Assumptions of normality for the distribution of grades in the criteria and exemplar groups were met, thus, inferential tests of Research Question 2(a) were parametric. Assumptions of normality of the distribution of grades and GPA in each of the self-feedback comment categories were not met, thus, non- parametric tests were used in the analysis of Research Question 2(b).

To test for any effect of the level of study on self-feedback comments produced by participants, we compared UG and PGT students on percentage of self-feedback comments in each category using Mann-Whitney U tests and found no significant difference between level of study in any of the four self-feedback comment categories (all ps >.29). To test for effects of the type of exemplar that students reviewed (peer work, or work produced by the teacher) Wilcoxon signed ranks tests compared each type of exemplar, i.e., student peer work, or work produced by the teacher, and found no significant differences between type of exemplar in the percentage of self-feedback comments in each of the 4 categories (all ps > 0.03, with α < = .006 after Bonferroni correction). A comparison of the effect of level of study (UG, PGT) on grades using an

independent t-test found no significant difference between UG and PGT participants (p = .28). Given the lack of an effect of level of study or type of exemplar, we collapsed across level of study and type of exemplar in all subsequent analyses. A breakdown of self-feedback comments according to level of study and type of exemplar can be found in the supplementary files.

Research Question 1: The Effect of Comparators (cohort) on Self-Feedback Comments

To answer research question 1, Mann Whitney U tests were used to compare the assessment criteria group (cohort 1) and exemplar group (cohort 2) on the number of self-feedback comments in each category (i) task non-elaboration; ii) task elaboration; iii) process –non-elaboration; iv) process elaboration. Mann-Whitney U tests (α < = .013 for multiple comparisons) showed that the percentage of task non-elaboration self-feedback comments was higher in the assessment criteria group [M = 70.27] than in the exemplar group [M = 29.09, U = 308.5, p <.001, r =.7], percentage of process non-elaboration comments was higher in the exemplar group [M = 32.36] than the assessment criteria group [M = 8.1, U = 351, p <.001, r = .68], and percentage of process elaboration comments was higher in the exemplar group [M = 20.52] than in the criteria group [M = 2.52, U = 492.5, p<.001, r = .64]. There were no significant differences between the assessment criteria and exemplar groups in task elaboration comments (p = .99). See Table 4 for summary statistics of percentage of self-feedback comments in each category for each comparison group (assessment criteria or exemplar), and Figure 2 for a graph of the means.

Table 4.

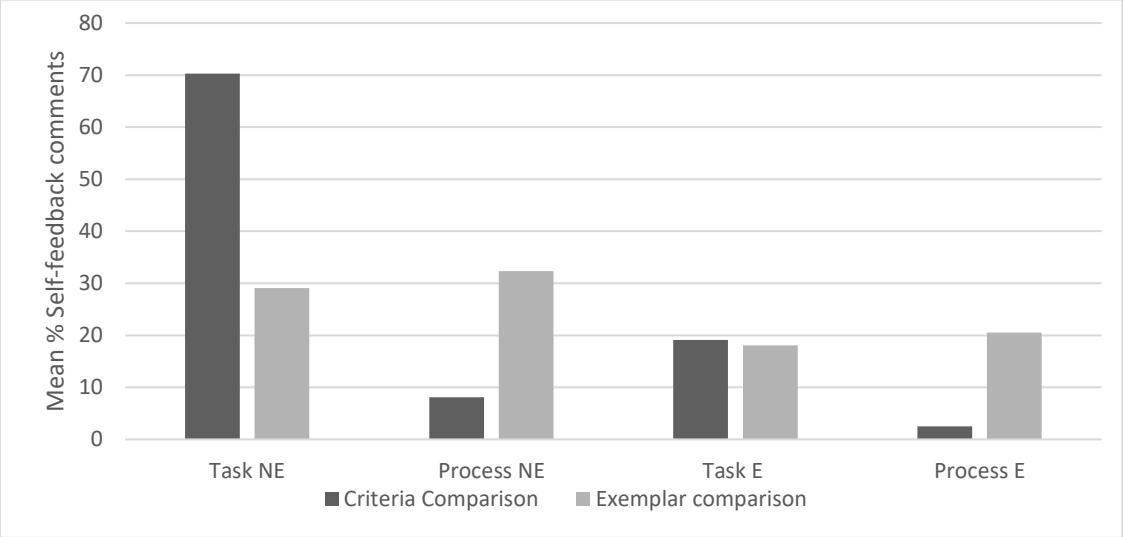
Percentage of participants' self-feedback comments in each of the four comment categories, according to comparison group.

Group	Comment Category	N	Min.	Max.	Mean	Standard Deviation
1: Comparison with criteria	Task Non-Elaboration	57	11.1	100	70.3	21.9
	Process Non-Elaboration	57	0	50	8.1	11.9
	Task-Elaboration	57	0	66.7	19.1	16.8
	Process-Elaboration	57	0	31.0	2.5	6.3
2: Comparison with exemplar	Task Non-Elaboration	56	0	87.5	29.1	20.6
	Process Non-Elaboration	56	0	68.9	32.4	15.7
	Task-Elaboration	56	0	55.2	18.0	13.5
	Process-Elaboration	56	0	69.1	20.5	17.7

Note: the minimum, maximum and mean values correspond to percentages of comments in each category.

Figure 2.

Graphic representation of the mean percentage of self-feedback comments in each of the four categories (Task Non-Elaboration, Process Non-Elaboration, Task Elaboration, Process Elaboration) according to comparison group (criteria or exemplar).



Research Question 2 (a): The Effect of Comparators on Grades

To answer research question 2 (a), an independent t- test compared the criteria and exemplar groups on their final assessment grade and showed that the exemplar groups' grade (M = 16.95, SD = 2.33) was significantly higher than the criteria groups' grade (M = 15.07, SD= 2.1), t (2,111), = 4.49, p <.001, d = .85). Descriptive statistics are provided in Table 4. To check for effect of academic year of study, an independent t-test compared the criteria and exemplar groups on their overall grade point average (GPA) and found no significant difference in their GPA (p = .113).

Table 4 Descriptive statistics of participants' mean grade for the final assessment and overall GPA for the academic year in each comparison group (assessment criteria, exemplar).

Group	N	Measure	Min	Max	Mean	Standard Deviation
Assessment Criteria	57	Assessment grade	10	21	15.1	2.3
		Overall GPA	13.4	20.1	16.6	1.4
Exemplar	56	Assessment grade	12	20	17.0	2.1
		Overall GPA	15.3	19.7	16.9	1.1

Research Question 2 (b): The Relationship between Self-Feedback Comments and Grade

To answer research question 2(b), the number of comments (expressed as a percentage) in each self-feedback comment category was correlated with participants' grades using Spearmans' Rho correlations. Initial correlations were run separately for the assessment criteria and exemplar groups, and the patterns of correlations did not differ substantially between cohorts (although there were differences in statistical significance due to the smaller sample size), so these were combined into a single data set for analysis. There was no relationship between task elaboration comments and grades, $r = -.037$, $p = .7$. There was a small significant positive correlation between grades and process non-elaboration comments, $r = .19$, $p = .04$, and between grades and process elaboration comments, $r = .25$, $p = .008$. There was a small significant negative relationship between task non-elaboration comments and grades, $r = -.2$, $p = .03$. Scatter plots are included in Figures 1, 2 and 3 in the supplementary files.

Discussion

Different Comparators Lead to Different Self-Feedback Comments (Research Question 1)

This study confirms prior research which shows that students can evaluate their own work and produce valuable self-feedback comments by making comparisons against information other than teacher comments (Nicol and McCallum., 2022; Nicol and Kushwah, 2023). However, it adds considerably to that research by providing evidence that different types of comparison information result in students producing different kinds of self-feedback. When students compared their own work against assessment criteria, they

produced more task non-elaboration feedback comments whereas comparisons against exemplars led them to produce more process comments. These process comments focused on how different aspects of the student's work inter-relate and how readers might interpret their work, a finding consistent with earlier research on exemplar comparisons (Nicol and Kushwah, 2023; Bouwer et al., 2018). The fact that there were no significant differences between the assessment criteria and exemplar conditions in the number of task elaboration comments, supports the conclusion that exemplar comparisons result in students shifting their focus from task to process feedback rather than simply producing more elaborate task-related comments. This shift has also been demonstrated by teachers, who deliver more process orientated comments, when providing feedback unconstrained by an assessment rubric (Dirkx et al., 2021).

Performance and Self-Feedback: a Complex Relationship (Research Question 2)

This study also compared the final assessment grade of the assessment criteria and exemplar groups and found that the exemplar group had a higher grade. This differs from earlier studies where an assessment rubric was superior to exemplars, when performance improvements were measured in a draft-redraft situation (Lipnevich et al, 2014; Lipnevich et al 2023) or measured by performance in a new task (Burnell et al, 2023). One likely reason for these differences across studies is that assessment criteria (the comparators used in this study) only provide information about performance goals, whereas a rubric (used in other studies) provides information about different levels of performance in relation to those goals (Wollenschläger et al., 2016).

There was a small, but significant, positive relationship between the process orientated comments and final grades and a small negative correlation between task orientated comments and grades. These findings align with previous studies showing that teacher-provided process comments lead to better student performance improvements (Wisniewski et al., 2020), and with Derham et al., (2021) who found that teacher comments containing corrective, or directive advice related to the task, negatively correlated with students' grades. However, in this study the process comments are self-generated by students not their teachers. This is arguably not only better in terms of students' learning but also for the development of their self-regulatory capacity.

In studies where feedback is provided by the teacher, the relationship between process comments and performance was, however, stronger than in the current study where students write their own feedback comments. The small effect size in our findings suggests that other possible moderating factors, in addition to

the nature of the comparators, are influencing students' performance. One known mediating factor is instructional prompts which have been shown to play a crucial role in what self-feedback students produce (Nicol, 2022). Another factor is the opportunities available to students to act on the feedback they produce. The extent to which students acted upon (or attempted to act on) their self-feedback at the time of its production is not known in this study, as this was not recorded, although it was evidenced in the final assessment. Future studies might involve having students act on their self-feedback at the time of its production as well as having them record the process (i.e. how they acted). This would not only allow more nuanced investigation of how students act on self-generated feedback but also of the role instructional prompts play in mediating the self-feedback process and its outcomes. Other researchers have also been calling for a better understanding of the internal mechanisms of feedback processing and for more diverse learning outcome measures (Panadero and Lipnevich 2022; Winstone & Nash, 2023).

Implications

Exemplars versus Assessment Criteria. This research has implications for teachers wishing to use exemplars and assessment criteria as a supplement or an alternative to giving students comments on their work. Exemplars enable students to generate self-feedback at both task *and* process level and they may be more beneficial as comparators if teachers need to make a choice. However, there are merits in using both rather than a single comparator. Assessment criteria may be more useful in helping students' tailor their work to the key requirements of the learning task whereas exemplars might be more useful in helping students take different and wider perspectives on their work and hence in helping them develop more elaborate knowledge structures (Bouwer, 2018). While Lipnevich et al (2014) found that using two comparators in the same task, rubrics and exemplars, reduced students' performance over one comparator, they attributed this to the cognitive load required to do two tasks in the same time frame. This could be addressed by staging the comparisons sequentially, one after the other.

Performance and learning: Making Self-Feedback Explicit and Prompts. Unpacking how learning and performance inter-relate is complex and more research to investigate this relationship is required. Burnell et al (2023) identify learning as long term performance, but examining this alone gives no insight into the inner processes of learning or what it is that causes long term performance gains. Learning is a construct that cannot be directly observed. This is why having students make their self-feedback productions visible in writing is valuable. It provides teachers with a window into students' thinking processes and hence their learning. The

more information teachers obtain about students' learning, the better they will be able to adapt their instructional strategies and to tailor their own feedback comments to students' needs.

However, it is important to note that having students make their internal feedback processes visible as self-feedback comments very likely changes the nature of those internal processes, making them more powerful (Nicol, 2021). Further, there is suggestive evidence that experience with this methodology results in students using it on their own without teacher direction (Nicol and Kushwah, 2023) which would ultimately reduce their reliance on the teacher for feedback. Making internal feedback processes visible is not just a valuable but overlooked educational feedback strategy but it is also affords a new way of researching feedback process, with many possibilities.

Limitations of Study

This study had some limitations. First, the investigatory context was a peer review task where students first produced comments on their peer's work before producing them on their own work. This meant that in the assessment criteria condition students used the assessment criteria twice (both when producing peer feedback and self-feedback) while those in the exemplar condition still used assessment criteria but only once (to provide peer feedback). This may have led those in the exemplar condition to also use criteria as a comparator, albeit implicitly, even though the prompts specifically instructed that group to compare their report against other exemplar reports. To address this potentially confounding factor, future research should be carried out in a non-peer review setting where students only produce self-feedback comments on their own work, by making criteria and exemplar comparisons, and not on that of their peers.

Another factor that may have influenced the exemplar comparison condition was variation in the quality of the exemplars used. While all students reviewed one identical high-quality exemplar and there were no differences in the type of self- feedback comments produced from peer exemplars and this high-quality exemplar, there was no control over the quality of the peer exemplars. For example, some students may have received one high and one low quality peer work for comparison while another may have received two low quality works. For a more fine-grained identification of how self-feedback comments evolve across multiple sequential comparisons, further exploration of the categories of self-feedback comments produced for each exemplar separately 1, 2 and 3 is required. Alternatively, the quality of all exemplars could be controlled by teachers selecting all of them from prior students' productions.

A third possible confounding factor in terms of students' final performance is that assessment criteria were used as a tool by teachers to grade the final assessment (rather than exemplars). By using criteria, teachers may themselves be paying more attention to task related aspects of the work rather than more holistic processes of students' work (Sadler, 2010). There is no easy solution to this problem apart from having teachers use a more holistic approach to assessing quality (see Biggs and Collis, 1982, for a possible way of achieving this).

Lastly, students in the criteria and comparison groups participated in the course in different academic years. Thus, it is possible that other factors in these two cohorts contributed to the difference in performance despite similar overall grades for the academic year in the two cohorts, and that participant groups had a similar balance of gender and nationality, and that the peer and self-feedback activities were identical in each academic year. Further examination of comparators within one year group and including a control group would however help identify within-group effects of self-feedback comments on performance.

Future research

The notion that different kinds of comparison information result in students producing different kinds of self-feedback comments is a new and unfolding area of research. While still at its early stage the focus to date has been limited to having students compare their work to information in assessment criteria, rubrics, exemplars, and comments. Future research is urgently needed to determine what self-feedback students might produce if they compared their work against information in resources beyond this limited set. Already researchers are suggesting that teachers could use this method to guide students to write self-feedback comments on their own critical and creative thinking (Nicol and Kushwah, 2023). For example, students who have produced an application of a theoretical model might compare that application against a different theoretical model and write self-feedback comments on the theory-application relationship.

Instructional prompts should also figure prominently in future research. Prompts can not only be used to focus students on learning processes but they can also be used to guide students to produce feedback on their own self-regulatory capacity. For example, students could be prompted to write self-feedback on how they would change their behaviour when they tackle future tasks, rather how they would improve performance in current tasks. Self-regulatory feedback is perceived as the most powerful type of feedback in the Hattie and Timperly's (2007) model. Having students self-generate it might accelerate the development of their own feedback agency. Finally, the emergence of generative Artificial Intelligence tools will make it easy

for teachers, or even students, to generate varied kinds of information for comparison (Yu, 2023) and to help formulate prompts. This will no doubt open many new possibilities for future practice and research.

Conclusion

In higher education, there is an inherent tension between teachers providing feedback comments to students and the wider educational goal that students become self-regulating and independent of their teachers (Allal, 2020; Nicol & Kushwah, 2023; Panadero & Lipnevich, 2022). Prior research suggests that we address this issue by having students take more agency in the commenting process by eliciting, requesting, and acting on the feedback that others provide, usually teachers (Carless and Boud, 2018). The findings of this study suggest a complementary approach to increasing students' feedback agency, namely, to have students compare their work against information in resources other than comments and to write their own feedback comments. When students write their own feedback comments it not only provides incontrovertible evidence of their own feedback agency, but it might also reduce teacher workload in commenting.

References

Allal, L. (2020). Assessment and the co-regulation of learning in the classroom. *Assessment in Education: Principles, Policy & Practice, 27*(4), 332-349. https://doi.org/10.1080/0969594X.2019.1609411

Andrade, H. L. (2019). A Critical Review of Research on Student Self-Assessment [Systematic Review]. *Frontiers in Education, 4*. https://www.frontiersin.org/articles/10.3389/feduc.2019.00087

Arts, J. G., Jaspers, M., & Joosten-ten Brinke, D. (2016). A case study on written comments as a form of feedback in teacher education: so much to gain. *European Journal of Teacher Education, 39*(2), 159-173. https://doi.org/10.1080/02619768.2015.1116513

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: the SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment and evaluation in higher education, 38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

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 [Original Research]. *Frontiers in Education, 3*. https://www.frontiersin.org/articles/10.3389/feduc.2018.00086

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics [Review]. *Frontiers in Education, 3*. https://doi.org/10.3389/feduc.2018.00022

Burnell, K., Pratt, K., Berg, D. A. G., & Smith, J. K. (2023). The influence of three approaches to feedback on L2 writing task improvement and subsequent learning. *Studies in Educational Evaluation, 78*, 101291. https://doi.org/https://doi.org/10.1016/j.stueduc.2023.101291

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

de Kleijn, R. A. M. (2023). Supporting student and teacher feedback literacy: an instructional model for student feedback processes. *Assessment & Evaluation in Higher Education, 48*(2), 186-200. https://doi.org/10.1080/02602938.2021.1967283

de Vreugd, L., Jansen, R., van Leeuwen, A., & van der Schaaf, M. (2023). The role of reference frames in learners' internal feedback generation with a learning analytics dashboard. *Studies in Educational Evaluation, 79*, 101303. https://doi.org/https://doi.org/10.1016/j.stueduc.2023.101303

Derham, C., Balloo, K., & Winstone, N. (2022). The focus, function and framing of feedback information: linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education, 47*(6), 896-909. https://doi.org/10.1080/02602938.2021.1969335

Dirkx, K., Joosten-ten Brinke, D., Arts, J., & van Diggelen, M. (2019). In-text and rubric-referenced feedback: Differences in focus, level, and function. *Active Learning in Higher Education, 22*(3), 189-201. https://doi.org/10.1177/1469787419855208

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Jensen, L. X., Bearman, M., & Boud, D. Characteristics of productive feedback encounters in online learning. *Teaching in Higher Education*, 1-15. https://doi.org/10.1080/13562517.2023.2213168

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559. http://www.jstor.org/stable/43575435

Lipnevich, A. A., & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions [Review]. *Frontiers in Education, 6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.720195

Lipnevich, A. A., Panadero, E., & Calistro, T. (2023). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of experimental psychology. Applied, 29*(1), 136-148. https://doi.org/10.1037/xap0000434

Lombard, M. (2010). *Intercoder Reliability*. Retrieved 06/06/2023 from http://matthewlombard.com/reliability/

Lui, A. M., & Andrade, H. L. (2022). The Next Black Box of Formative Assessment: A Model of the Internal Mechanisms of Feedback Processing [Hypothesis and Theory]. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.751548

Maguire, M. D., B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education, 9*(3), 3351-33514. https://ojs.aishe.org/index.php/aishe-j/article/view/335

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education, 45*(4), 527-540.

https://doi.org/10.1080/02602938.2019.1667955

Narciss, S., Prescher, C., Khalifah, L., & Körndle, H. (2022). Providing external feedback and prompting the generation of internal feedback fosters achievement, strategies and motivation in concept learning. *Learning and Instruction, 82*, 101658. https://doi.org/https://doi.org/10.1016/j.learninstruc.2022.101658

Nicol, D. (2021). The power of internal feedback: exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756-778. https://doi.org/10.1080/02602938.2020.1823314

Nicol., D. (2022). Turning Active Learning into Active Feedback: Introductory Guide from Active feedback toolkit. National Teaching Repository. https://figshare.edgehill.ac.uk/articles/educational_resource/Active_Feedback/19929290 DOI: 10.25416/NTR.19929290.v3

Nicol, D., & Kushwah, L. Shifting feedback agency to students by having them write their own feedback comments. *Assessment & Evaluation in Higher Education*, 1-21. https://doi.org/10.1080/02602938.2023.2265080

Nicol, D., & McCallum, S. (2022). Making internal feedback explicit: exploiting the multiple comparisons that occur during peer review. *Assessment & Evaluation in Higher Education, 47*(3), 424-443. https://doi.org/10.1080/02602938.2021.1924620

Nicol, D., & Selvaretnam, G. (2022). Making internal feedback explicit: harnessing the comparisons students make during two-stage exams. *Assessment & Evaluation in Higher Education, 47*(4), 507-522. https://doi.org/10.1080/02602938.2021.1934653

O'Donovan, B., Rust, C., & Price, M. (2016). A scholarly approach to solving the feedback dilemma in practice. *Assessment & Evaluation in Higher Education, 41*(6), 938-949. https://doi.org/10.1080/02602938.2015.1052774

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

Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning Self-Assessment into Self-Feedback. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners* (pp. 147-163). Springer International Publishing.

https://doi.org/10.1007/978-3-030-25112-3_9

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, 100416. https://doi.org/https://doi.org/10.1016/j.edurev.2021.100416

Price, D., Smith, J. K., & Berg, D. A. G. (2017). Personalised feedback and annotated exemplars in the writing classroom: An experimental study in situ. *Assessment Matters, 11*, 122-144. https://doi.org/10.18296/am.0027

Purchase, H., & Hamer, J. (2018). Perspectives on peer-review: eight years of Aropä. *Assessment & Evaluation in Higher Education, 43*(3), 473-487. https://doi.org/10.1080/02602938.2017.1359819

Rust, C., Price, M., & O'Donovan, B. (2003). Improving Students' Learning by Developing their Understanding of Assessment Criteria and Processes. *Assessment and evaluation in higher education, 28*(2), 147-164. https://doi.org/10.1080/02602930301671

Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550. https://doi.org/10.1080/02602930903541015

Smith, J. K., A.A. Lipnevich & T.R. Guskey. (2023). *Instructional Feedback: The Power, the Promise, the Practice.* Corwin.

Soper, D. (2024). *Free Statistics Calculators*. Retrieved 01/12/ 2023 from https://www.danielsoper.com/statcalc/default.aspx

Taylor, B., Kisby, F., & Reedy, A. (2024). Rubrics in higher education: an exploration of undergraduate students' understanding and perspectives. *Assessment & Evaluation in Higher Education*, 1-11. https://doi.org/10.1080/02602938.2023.2299330

To, J., Panadero, E., & Carless, D. (2022). A systematic review of the educational uses and effects of exemplars. *Assessment & Evaluation in Higher Education, 47*(8), 1167-1182. https://doi.org/10.1080/02602938.2021.2011134

Tomazin, L., Lipnevich, A. A., & Lopera-Oquendo, C. (2023). Teacher feedback vs. annotated exemplars: Examining the effects on middle school students' writing performance. *Studies in Educational Evaluation, 78*, 101262. https://doi.org/https://doi.org/10.1016/j.stueduc.2023.101262

Underwood, J. S., and Tregidgo, A. P. (2006). Improving student writing through effective feedback: best practices and recommendations. *Journal of Teaching Writing, 22*, 73-98.

Van der Kleij, F. M., Adie, L. E., & Cumming, J. J. (2019). A meta-review of the student role in feedback. *International journal of educational research, 98*, 303-323. https://doi.org/10.1016/j.ijer.2019.09.005

Wiliam, D. (2018). Feedback: At the heart of—but definitely not all of—Formative Assessment. In Lipnevich, A. A., & Smith, J. K. (Eds.), *The Cambridge handbook of instructional feedback* (pp3-28). Cambridge University Press. https://doi.org/10.1017/9781316832134

Winstone, N. E., & Nash, R. A. (2023). Toward a cohesive psychological science of effective feedback. *Educational Psychologist, 58*(3), 111-129. https://doi.org/10.1080/00461520.2023.2224444

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

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology, 44-45*, 1-11. https://doi.org/https://doi.org/10.1016/j.cedpsych.2015.11.003

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching [Opinion]. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1181712

Supplementary files and the data used in the analysis can be found on OSF

https://osf.io/64auv/?view_only=6a4d49e6bf2c424592dde3a72ac2b2c4