Studies in Educational Evaluation 78 (2023) 101262

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Teacher feedback vs. annotated exemplars: Examining the effects on middle school students' writing performance

Ligia Tomazin [a,1,*], Anastasiya A. Lipnevich [b,2], Carolina Lopera-Oquendo [a,3]

[a] The Graduate Center, City University of New York, USA
[b] Queens College and the Graduate Center, City University of New York, USA

ARTICLE INFO

Keywords:
Self-feedback
Internal feedback
Corrective feedback
Exemplars
Middle-school
Writing task

ABSTRACT

The current study investigated students' improvement on a writing task following the use of annotated exemplars, teacher comments, and the combination of both approaches. A sample of 94 middle school students (age $M = 12.42$, $SD = 0.96$) from a private school in Brazil was randomly assigned to one of three feedback conditions: annotated exemplars, teacher comments, and both annotated exemplars and teacher comments. Participants were asked to write an essay and then revise it by using teacher comments or annotated exemplars (or both). Results showed improvements in students' writing from first to second draft, but no statistically significant differences among the groups were found. Further, girls scored higher than boys in both the first and final drafts irrespective of the feedback condition. These results show the promise of annotated exemplars in facilitating students' improvement on a writing task through effective self-feedback generation while significantly reducing teachers' time investment.

The efficacy and importance of feedback for students' improvement of performance and learning is well recognized (Kluger & DeNisi, 1996; Lipnevich & Panadero, 2021). However, it is noteworthy that the effectiveness of feedback is strongly dependent on students' active engagement and processing of feedback (Lipnevich & Smith, 2022; Winstone et al., 2017; Nicol, 2020). Researchers have argued that after internalizing feedback from external sources, students convert this information into self- or inner feedback, which becomes the driver of changes in learning and performance (Lipnevich & Smith, 2022; Narciss et al., 2022; Panadero et al., 2019). There is a multitude of ways in which feedback can be delivered to students, and teacher comments have been traditionally regarded as the gold standard of feedback provision (Lipnevich & Smith, 2009b; Nicol, 2020). However, there is a general consensus among researchers and practitioners alike that providing effective individualized feedback to students is not a simple task for both experienced and early career teachers (Carless et al., 2011; Mañez et al., in press). It is also a very time-consuming undertaking for all instructors (King et al., 2008; Lipnevich et al., 2022; Price et al., 2010). To address this problem, educators have begun exploring alternative ways of feedback provision that would be efficient, effective, and will help students to engage in successful self-feedback generation and thus enhance their performance (Lipnevich et al., 2014; 2022).

For example, rubrics and exemplars, when used as feedback – that is, after students produce their initial assignment drafts – have shown promising results in terms of quality of students' writing improvement (e.g., Lipnevich et al., 2014; 2022; Nicol & McCallum, 2022). Although somewhat counter-intuitive, the use of these tools after students' work had been submitted fits into the current definitions of feedback. So, Lipnevich and Smith (2022) broadly defined feedback as any information related to students' performance that offered opportunities for improvement. Therefore, any instructional tool that is presented after initial drafts are submitted and that can help students to improve would be regarded as feedback.

In considering mechanisms of feedback processing that could explain the effectiveness of instructional tools delivered after students' drafts have been submitted, Nicol (2020) suggested that students rely on comparison processes between their current performances and external sources to generate self- (or inner) feedback. Exploring the effectiveness of self-feedback generation as students engage with different instructional tools (such as exemplars and rubrics) and tasks (such as peer

* Correspondence to: Department of Educational Psychology, the Graduate Center, The City University of New York, 365 5th Avenue, New York, NY 10016, USA.
E-mail address: ligiatfm@gmail.com (L. Tomazin).
[1] https://orcid.org/0000-0002-9753-3246
[2] https://orcid.org/0000-0003-0190-8689
[3] https://orcid.org/0000-0002-9355-5843

https://doi.org/10.1016/j.stueduc.2023.101262
Received 6 October 2022; Received in revised form 29 January 2023; Accepted 11 April 2023

review), has become a thriving area of research (e.g., Lipnevich et al., 2014; 2022; Nicol & McCallum, 2022).

Considering that helping students of all levels to become less dependent on teacher comments and more capable of generating effective self-feedback is one of our collective goals, further disentangling the effects of exemplars and other tools on student performance is of great value to the field. Furthermore, understanding the relative effectiveness of such tools compared to teacher comments is critical. Teachers often feel compelled to spend exorbitant amounts of time on individualized feedback delivery, and freeing time to prepare new content or diversify instructional activities would be beneficial for all involved, should other, more scalable feedback approaches be deemed as effective (Nicol & McCallum, 2022). Hence, the study reported herein explored the effects of teacher comments and annotated exemplars, when used as feedback, on middle school students' revision of their writing assignments.

1. Exemplars and the development of writing

Writing is a complex skill that evolves over time (Harris & McKeown, 2022). The importance of developing writing skills in students cannot be disputed as it predicts academic success across educational levels (Graham & Perin, 2007). After all, there is hardly an academic domain that does not rely on writing. Due to its complex nature, developing students' writing is not a straightforward task. Researchers have consistently concluded that revisions represented a critical step in writing improvement (Holtz & Daly, 2021; MacArthur, 2018). Effective revision processes encourage learners to revisit their work and make judgments that lead to changes in the quality, clarity, and cohesion of the written discourse, in frequency of mechanical errors (such as spelling and punctuation), in arguments, and communication with the audience, among others (Graham, 2018; Graham & MacArthur, 1988; MacArthur, 2018). Further, frequent and targeted feedback, in the context of revisions, has been deemed to be one of the key instructional tools that can promote improvement. For example, in a meta-analysis of the effects of formative assessment on writing, Graham et al. (2015) found a positive effect of feedback from adults, peers, self, and computers on the quality of children's writing. At the same time, in a study concerning teaching writing to middle school students, more than half of the participating teachers said that they restricted the amount of writing and revision activities included in their lessons because of the excessive time they needed to assess those tasks (Graham et al., 2013). That is, evaluation of student drafts, feedback provision, and opportunities to revise all come at a high price – teachers' time. Thus, finding ways to continue developing student writing skills while keeping teachers' time investment manageable is a problem that may be solved with the help of exemplars and other instructional tools.

So, what are exemplars? In writing, exemplars are templates specifically selected to indicate a desired level of quality or proficiency (Sadler, 1987). Students' preference for having access to exemplars has been repeatedly reported in research (Bell et al., 2013; Handley & Williams, 2011; Yang & Zhang, 2010), although the contribution of this tool to student writing improvement remains rather unclear. Furthermore, it is important to distinguish the different roles exemplars may fulfill, which vary according to the timing they are introduced to students. For example, they may be used to illustrate and clarify assessment standards (Bell et al., 2013; Broadbent et al., 2018) or to strengthen students' comprehension of the assessment criteria (Handley & Williams, 2011; Hendry et al., 2016). In those cases, exemplars are presented to students before they engage in a task. However, the effects of exemplars on student writing in this scenario are contradictory. Handley & Williams (2011) conducted a quasi-experimental study to explore the effect of exemplars on student writing. Despite high engagement of students with exemplars, they found no improvement in the quality of students' performance when compared to the performance of students from former cohorts who did not have access to exemplars. Conversely, Rust et al. (2003) revealed better quality of coursework for business students who participated in a tutor-led intervention where they engaged in active marking, grading, and discussion of exemplars of assignments. Additionally, in a similar intervention involving discussion and marking of exemplars, Hendry et al. (2016) found no effects of engagement in exemplar marking on student performance on an assignment.

An alternative way of using exemplars is to present them to students after they have completed an initial draft of their assignments. That is, using exemplars as feedback (Lipnevich et al., 2014). At this stage, students can compare what they have accomplished to an example of what is expected from them and generate ideas for how to improve their work, thus using exemplars to generate self-feedback. For instance, Lipnevich et al. (2014) found that having access to exemplars of various levels of task performance *after* the participants had produced an initial draft positively impacted the quality of psychology students' revision of a research proposal. Extending these findings, Lipnevich et al. (2022) reported positive effects of exemplars on high school students' writing performance, especially after receiving training on how to use these resources. Similarly, Yang and Zhang (2010) showed evidence of improvements in the quality of ten English as Second Language students' writing assignments after these participants compared their initial draft to a version of it written by a native speaker.

In Lipnevich et al. (2014, 2022), students in the exemplars group significantly improved their performance, but not as much as students in the rubrics group. One of the explanations that the researchers proposed was that criteria of success in rubrics were explicit, and hence, did not require as intensive a processing. To this end, Nicol (2020) emphasized the importance of making the comparison process explicit to enhance its power for the quality of self-feedback and, consequently, performance. Thus, providing annotations in the exemplars to clearly direct learners attention to specific criteria has the potential to facilitate the revision process.

Overall, the interest in exploring the potential value of exemplars has been growing (To & Carless, 2016), but, to our knowledge, there is very limited research comparing exemplars' effectiveness to that of teacher comments. Price et al. (2017) is one such study and it was designed to examine the relative effectiveness of teacher comments and exemplars. In this experimental study, the authors collected four separate sets of data from New Zealand students enrolled in grades 9 and 10 ($n = 40$). In the quantitative portion of this study, students participated in two cycles of writing, which encompassed writing an initial draft, receiving feedback, and revising their work. After initial drafts were completed, participants were randomly assigned to a feedback condition and received either personalized feedback or annotated exemplars in the first cycle, whereas in the second cycle, the condition switched. In addition to receiving either teacher comments or exemplars, students had the opportunity to discuss feedback with the instructor. Participants in the teacher comments group met with the teacher to go over their feedback (each student had approximately 8 min of one-on-one meeting time with the teacher), whereas students in the annotated exemplars condition participated in a group discussion with the teacher (nearly 30 min per group). Despite students' preference for teacher comments, results showed no statistical difference in the improvement of performance between students who received individualized comments and those who received annotated exemplars. However, because both groups engaged in discussions, it was impossible to disentangle the effect of feedback itself from its discussions (group or individual) on writing revisions. In other words, the design of the study prevents us from concluding how much those discussions impacted students' generation of self-feedback and their subsequent performance. Furthermore, with the exception of Price et al. (2017), studies investigating the efficacy of exemplars as a form of standardized feedback have been focused on samples of more experienced learners (Lipnevich et al., 2014; Lipnevich et al., 2022; Yang & Zhang, 2010).

When it comes to individual differences and student responses to

feedback, studies exploring how different types of standardized feedback may impact learners depending on their genders and grade levels are lacking. Andrade and Boulay (2003) found that 7th and 8th grade girls in the treatment condition benefited from self-assessment practices using rubrics in writing fictional essays in history whereas there was little to no significant relationship between those practices and boys' writing. Interestingly, neither boys nor girls benefited from the treatment condition in their performance on a literature essay. In contrast, Huang and Wilson (2021) found that after receiving automated feedback on their writing, boys in grade 4 and 5 improved at a slightly faster pace than girls despite having produced poorer first drafts. These nuances in how different formats of feedback might differentially impact students of different genders and grades need to be further explored, so that practitioners can provide adequate feedback to their students. Therefore, to a) downward extend findings of earlier studies on exemplars, to b) compare students' improvement on a writing task based on either exemplars of effective work, teacher comments, or a combination of the two, and to c) explore differences in gender and grade levels when receiving those forms of feedback, we conducted the current study.

2. The current study

Our purpose with this experimental study was to investigate middle school students' improvement on a writing task following the use of annotated exemplars and teacher comments (or both). Having a greater understanding of how middle school students can generate self-feedback based on annotated exemplars and teacher comments has both theoretical and practical implications for the development of writing. Investigations involving primary and middle school learners have been scarce, so our results will shed some light on students' ability to revise their work based on annotated exemplars, teacher comments, or both. Furthermore, there are very few studies exploring gender differences in student performance as they engage in self-assessment (Rust et al., 2013; Huang & Wilson, 2021) but, to our knowledge, there is no research investigating gender differences in student performance when exemplars are used as a source of feedback. Hence, we attempted to answer the following research questions:

1. Are there differences in student improvement on a writing task depending on the type of feedback they receive (individualized teacher comments and/or annotated exemplars)?
2. Are there gender differences in student improvement depending on the feedback condition?
3. Do students of different grade levels differentially improve depending on the feedback they receive?

3. Method

3.1. Participants

Participants were 94 middle school students[4] (32 6th graders, 35 7th graders, and 27 8th graders) from a private school in the suburbs of São Paulo, Brazil. Their ages ranged from 11 to 14 years ($M = 12.42$, $SD = 0.96$) and 53.2% of the participants were female.

3.2. Procedure

Fig. 1 depicts the sequence of study procedures. Students were invited to write a persuasive essay. This task was part of the school's planned activities and therefore was also completed by the non-participating peers. In order to allow their work to be included in the study, students had to return their parents' signed consent form and also had to assent to study participation. Participation was voluntary and approximately 30% of the middle school students from the selected school returned their documentation in a timely manner and chose to join the study.

During the pre-study session teachers introduced and explained the assignment. Students were then given a period of seven days to write their essays and submit their first draft through a virtual platform regularly used by these students (Time 1, Session 1). Each student was randomly assigned to one of the three feedback conditions using an electronic randomization app (Time 1, Session 2). Therefore, 31 students were assigned to the individualized teacher feedback condition, whereas 33 and 30 students were assigned into the exemplars and combined conditions, respectively. Table S1 (Supplementary Materials) shows the distribution of participants by condition, grade, and gender.

Teacher comments: In this condition, students received comments on their work. Students' mistakes in spelling, punctuation, style, structure, and content were pointed out and suggestions were provided. The necessary time to provide this type of feedback on each essay ranged from 8 to 15 min.

Exemplars: Students in this condition received two well-written exemplars on the same topic. Offering middle school students, a single exemplar could have potentially led them to closely imitate it. By offering them two high quality exemplars we hoped to encourage students to see beyond the surface structure of the essay. Those models were annotated, and commentaries were added to make certain aspects of the structure and argumentation explicit. That is, criteria of optimal performance were explicated. The language used in the comments prompted students to compare the exemplars to their own work and to decide whether the criteria of successful performance had been met. The first author and teachers worked together to identify frequent mistakes found in students' writing in order to direct the annotations on the exemplars. The time used to annotate each exemplar ranged from 26 to 35 min. The annotated exemplars can be found in Supplementary Materials.

Teacher comments and annotated exemplars (combined): Students in this condition, in addition to receiving comments on their work, were also given access to the two annotated exemplars.

Unfortunately, we were not allowed to use a control group. A similar study conducted by Lipnevich and Smith (2009) found that receiving no feedback but an opportunity to revise an essay showed no improvement in students' scores (Hedges $g = 0.012$, ns), whereas students with detailed feedback improved strongly ($g = 1.23$, $p < 0.001$). It is beyond the scope of this study to investigate the effectiveness of instructional feedback. Instead, our major goal was to investigate the potential of annotated exemplars as a form of feedback and whether the effects of exemplars could be comparable to those of teachers' comments.

After essays were assessed, instructors posted comments or offered annotated exemplars (or both) on the virtual platform, and students were encouraged to use the comments and/or exemplars to revise their work again. One week later, students submitted their revised essays through the virtual platform (Time 2). Students' writing was again evaluated and graded. All participants received access to their grades pre- and post-feedback after the end of the writing cycle.

All data collection was conducted virtually and there was no personal contact between researchers and students.

3.3. Measures

3.3.1. Grades

Following the school assessment standards, draft and final essays were graded on a continuous scale from 0 to 10. One member of the research team graded both the initial draft and the final assignment using a detailed rubric developed for the evaluation of this assignment. Additionally, a trained teacher graded 22% of the essays using the same criteria to establish inter-rater reliability. Inter-rater consistency on the

[4] The sample size was determined through a power analysis conducted in G*Power 3.1 (Faul et al., 2009). With an effect size of 0.4, a power of 0.90, a significance level of 0.05, and a correlation among repeated measures of 0.8, the recommended sample size was 75 participants.

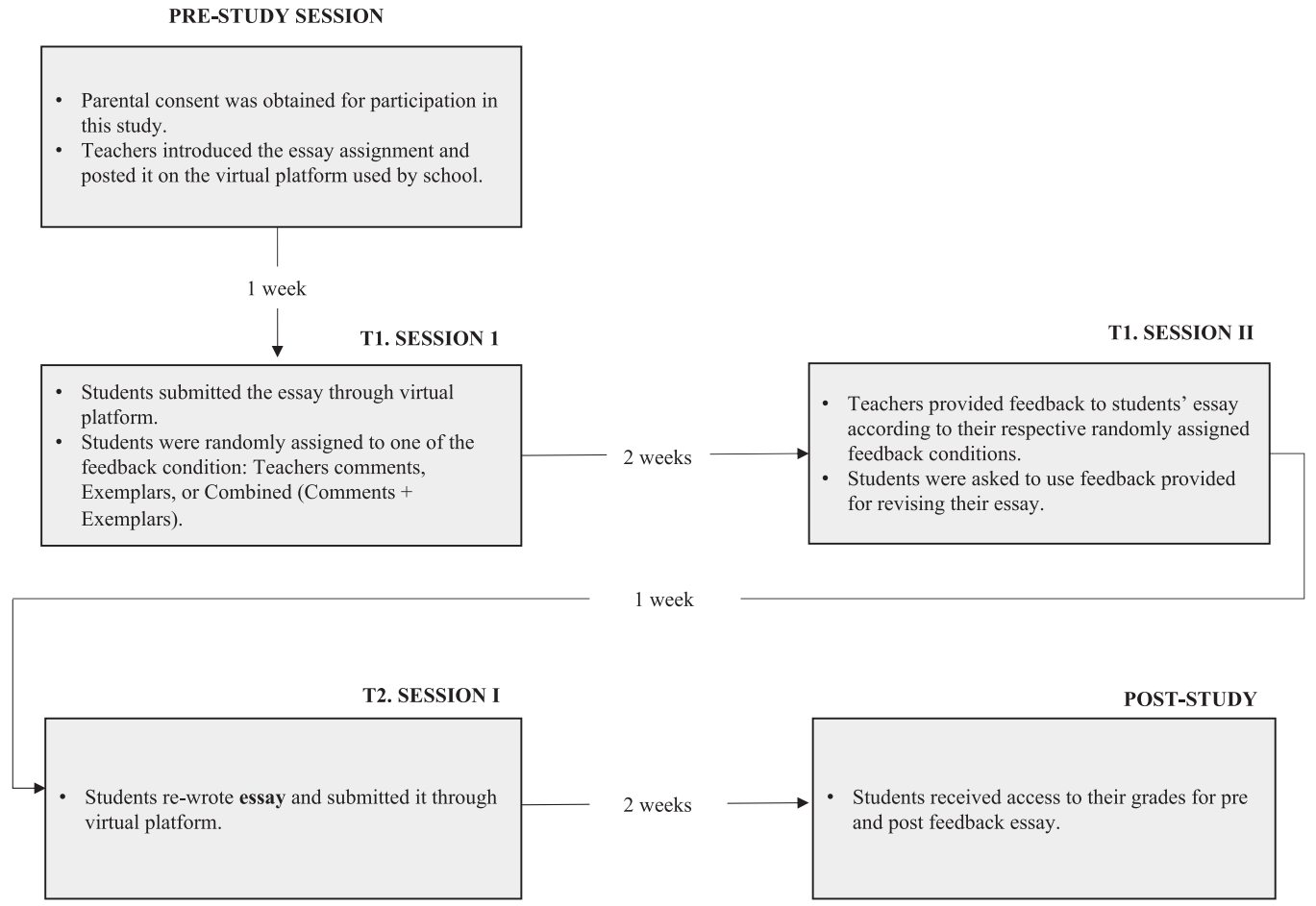

Fig. 1. Procedure flowchart.

subsample was calculated using the intraclass correlation coefficient (ICC). An ICC of .783 showed a good reliability between the two raters (95% CI [.462 < ICC <.914], $p < .001$). Additionally, the Pearson correlation coefficient between graders' scores was .852. Both graders were blind to the students' condition. Table S2 (Supplementary Materials) shows descriptive information for grades. For the total sample, the score for draft essays ranged between 5.8 and 9.4 points (M= 7.49, SD= 0.97), whereas the scores for final grades ranged from 6.2 to 10.0 (M= 8.22, SD= 1,03). Additionally, skewness and kurtosis values were in the − 2 and + 2 range, which is considered acceptable (George & Mallery, 2010).

3.3.2. Demographic information

Information about students' gender and grade level was collected through the school's administrative records.

3.4. Analytic plan

Descriptive statistics were computed. To answer our research questions, we conducted a four-way mixed ANOVA with one repeated or within-subject factor (essay score in time 1 and 2) and three between-subject factors (type of feedback, gender, and grade). Assumptions about outliers, normality, and homogeneity of variances were tested. Post-hoc analysis, using Benjamini-Hochberg as a method to adjust p-values, was also conducted. All analyses were carried out using R software version 4.1.2 (R Core Team, 2021).

4. Results

4.1. Preliminary analysis: descriptive information

Descriptive statistics of students' grades according to each feedback condition are presented in Table 1. Effect sizes ranged from $d = 1.2$ for Teacher Comments and Comments and Exemplars (Combined) conditions to $d = 1.89$ for the Exemplars-only conditions, indicating substantial differences (Lakens, 2013). Additional information about grades distribution by condition is presented in Table S2 (Supplementary Materials).

4.2. Effect of feedback condition, grade, gender, and time on writing scores

A four-way ANOVA (Table 2), with the type of feedback, gender, and grade level as between-subject factors and students' score in time 1 (draft essay) and time 2 (final essay) as within-subject variable was conducted. Homogeneity of variance and normality assumptions were

Table 1
Descriptive Statistics of Students' Grades across Feedback Condition.

Feedback Condition	N	Grade Draft (Time 1)		Grade Final (Time 2)		Cohen's d
		M	SD	M	SD	
Teacher Comments	31	7.64	0.99	8.33	1.09	1.21
Exemplars	33	7.57	1.02	8.35	1.03	1.89
Combined	30	7.23	0.88	7.97	0.94	1.29
Total	94	7.49	0.97	8.22	1.02	1.43

Table 2
Four-way Mixed ANOVA Results.

Measure	df	F	p	Partial η^2
Within-subject effects				
Time	1	193.180	.000*	.718
Grade:Time	2	0.987	.377	.025
Gender:Time	1	7.032	.010*	.085
Condition:Time	2	0.029	.971	.001
Gender:Grade:Time	2	0.633	.534	.016
Condition:Grade:Time	4	1.898	.119	.091
Condition:Gender:Time	2	0.305	.738	.008
Condition:Gender:Grade:Time	4	0.966	.431	.048
Between-subject effects				
Gender	1	9.017	.004*	.106
Grade	2	2.120	.127	.053
Condition	2	0.481	.620	.013
Gender:Grade	2	0.127	.881	.003
Condition:Grade	4	0.620	.650	.032
Condition:Gender	2	1.253	.292	.032
Condition:Gender:Grade	4	0.474	.755	.024

* $p < 0.05$

tested. The Shapiro-Wilk test was computed for each combination of factor levels for checking normality. Tests indicated that differences in scores were normally distributed ($p > .05$) except for three groups (boys in grade six with exemplars condition in time 1 ($p = .006$), and boys in seventh grade in the teacher comments condition ($p < .001$) and the exemplars conditions ($p = .030$ in time 2). QQ-plot for each cell of the analysis (Fig. S1, Supplementary Materials) also indicated that all points fell approximately along the reference line. Additionally, Levene's test for homogeneity of variance was not significant (Time 1, $p = .836$; Time 2, $p = .957$), so we assumed normality and homogeneity of the residual variances for all groups.

Results showed that there were no statistically significant differences in student scores among the three experimental conditions ($F(2, 76) = 0.481$, $p = .620$) and grade levels ($F(2, 76) = 2.120$, $p = .127$). Also, the interaction effect between condition and time was not statistically significant ($F(2, 76) = 0.029$, $p = .971$), suggesting that were not differences in the initial (Time 1) and final (Time 2) performance of students across the three experimental groups. Hence, the randomization across conditions was effective. The main effect for gender ($F(1, 76) = 9.017$, $p = .004$, *partial* $\eta^2 = 0.106$) and time ($F(1, 76) = 196.18$, $p < .001$, *partial* $\eta^2 = 0.718$) on students' grades were statistically significant. Additionally, the two-way interaction between gender and time was also statistically significant ($F(1, 76) = 7.032$, $p = .010$, *partial* $\eta^2 = 0.085$). For interactions and simple effects, a pairwise comparison test was used with Benjamini-Hochberg adjustment.

Pairwise comparisons indicated that writing task scores before receiving feedback (draft essay, time 1) ($M = 7.49$; $SD = 0.97$) were statistically different from the scores after feedback (final essay, time 2) ($M = 8.22$; $SD = 1.03$) ($p.adj < .001$). Additionally, there was a statistically significant effect of student gender on writing scores for both scores before ($p.adj = .004$) and after ($p.adj < .001$) feedback. That is, boys in time 1 (M=7.182, SD=0.763) and time 2 (M=7.795, SD=0.850) received lower scores than girls (time 1:M=7.752, SD= 1.065 and time 2:M=8.596, SD= 1.031), respectively). However, both groups of students showed a significant increase in their writing performance after feedback ($p.adj < .001$) (Fig. 2). Moreover, girls and boys alike had a statistically significant improvement in their final scores in comparison to draft essay across all feedback conditions ($p.adj < .05$) (Table S3, Supplementary Materials), except for boys in the teacher comments condition ($p.adj = .181$) (Fig. 3).

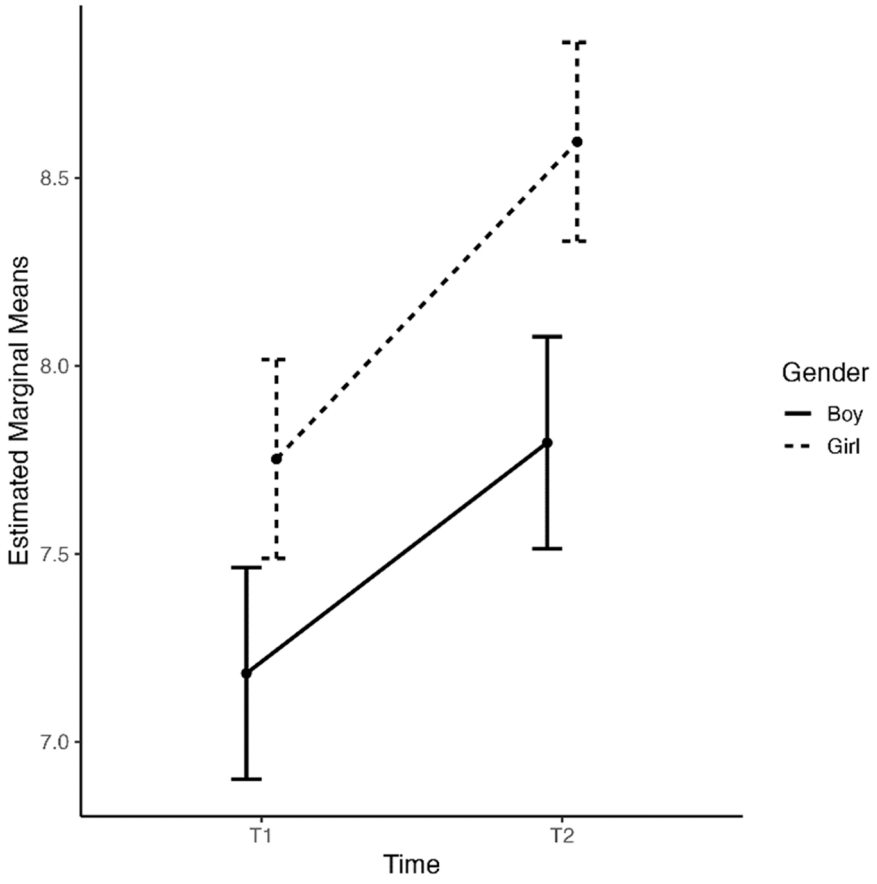

Fig. 2. Estimated Marginal Effects by Gender and Time on Writing Task.

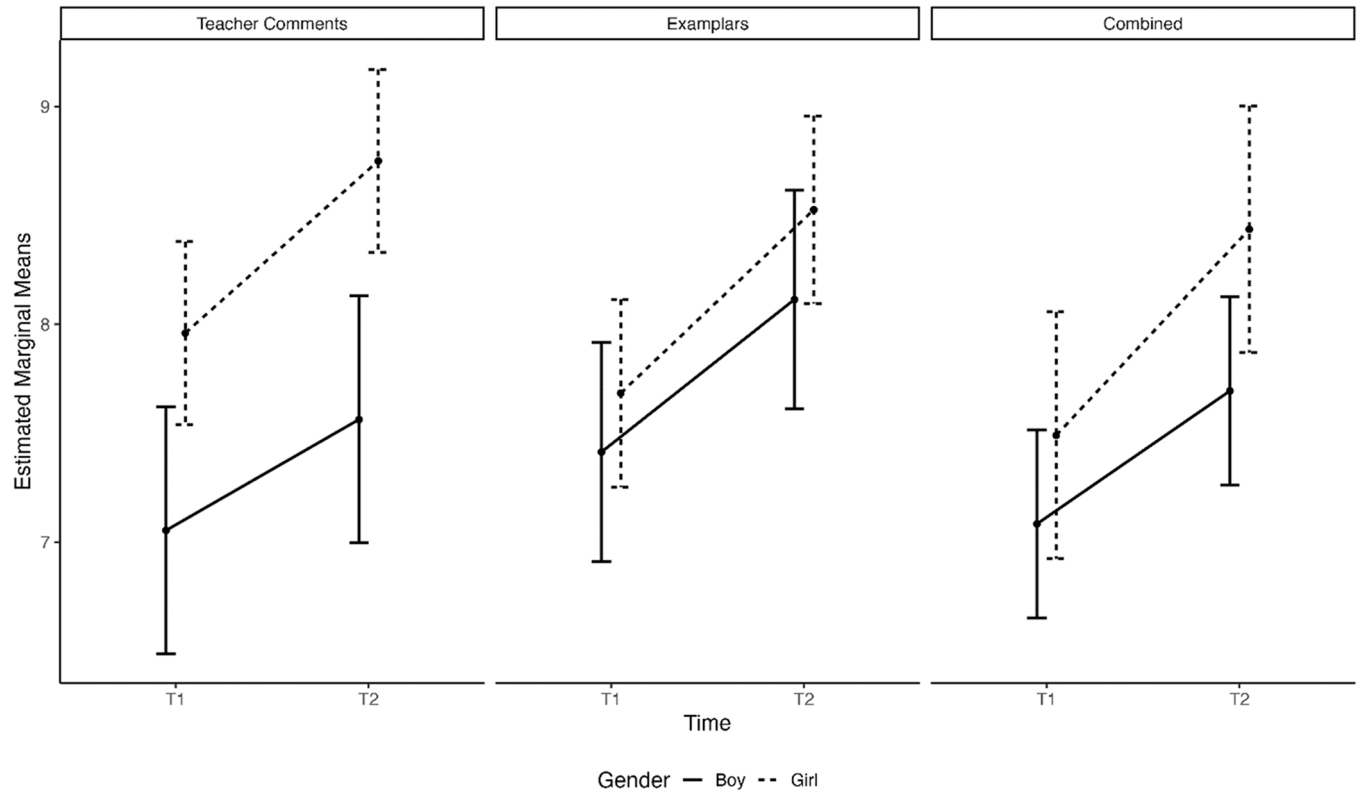

Fig. 3. Estimated Marginal Effects by Condition, Gender, and Time on Writing Task.

5. Discussion

In this study we compared the effects of annotated exemplars, individualized teacher comments, and the combination of both teacher comments and exemplars on middle-school students' revision of their writing assignments. Both teacher comments and annotated exemplars were presented after students submitted their initial drafts and, hence, represented a form of feedback. Our results showed that students in all three conditions significantly improved their performances from first to second draft. However, there were no statistically significant differences among students who received comments from the teachers, those who received annotated exemplars of excellent essays, and those who had access to both types of feedback.

These findings may suggest that even without explicit guidance on how to improve their work that is offered through teacher comments, annotated exemplars prompted students to compare their work to exemplars of optimal performance and thus helped students to generate effective self-feedback and improve their work. Even though teacher comments are widely used to facilitate revisions, standardized feedback in the form of exemplars led to comparable results in our investigation. Our results are in line with findings of Price et al. (2017), where no statistically significant differences were found between annotated exemplars and personalized feedback on a writing task of New Zealand secondary school students. However, in contrast to Price's et al. (2017) study, where students were also involved in the discussion of teacher comments and exemplars, our participants' generation of self-feedback and improvement in performance can be attributed exclusively to students' engagement with annotated exemplars and teacher comments, as they did not take part in any discussion of feedback.

Moreover, we explored the potential effects of students' gender and grade level on their performance improvement. We found that female students showed higher levels of performance overall and a greater improvement, irrespective of the feedback condition to which they had been assigned. This finding is consistent with those of earlier studies, showing that girls outperformed boys in language courses and tasks (Voyer & Voyer, 2014). Our study also suggested that although the net improvement in student performance after an initial draft was higher for girls, both genders equally benefited from their engagement with exemplars, teacher comments, or the combination of these two approaches. In a somewhat different but related context, Huang and Wilson (2021) found that boys had lower scores than girls on their initial drafts, even after controlling for language proficiency, race, and socio-economic status, but increased at a slightly faster rate than girls when using automated feedback. These results emphasized the importance of closer examination of gender differences in the use of feedback in general and exemplars in particular.

Although our results should be interpreted with caution, they are very optimistic for two reasons. First, when given the opportunity to revise, students significantly improved their writing. Second, teacher comments that are regarded as the gold standard of feedback and that also require tremendous investment of time, have worked just as well as annotated exemplars in our study. Let us consider these two contentions more closely. Students need opportunities to exercise and develop their writing skills, and many studies have shown benefits of revisions (Graham & Perin, 2007; Graham, 2018; Graham et al., 2021). After all, even the best writers need to step away and re-engage with their work, and for novice writers this opportunity should be a given (Graham, 2018). Studies also reveal significant improvements in student writing, depending on the feedback they receive (Graham et al., 2015; Lipnevich & Smith, 2009). The key role of feedback as a critical tool in the development of writing has traditionally been interpreted through the use of comments from teachers. However, this practice is highly time consuming for practitioners who often restrain the amount of required practice for students due to their personal time investment associated with provision of good quality feedback (Graham et al., 2013). Researchers have argued (Lipnevich & Smith, 2022; Nicol, 2020) that an alternative to providing individualized feedback would be presenting students with detailed standardized feedback that will allow students to

engage in self-feedback generation and improve their work. Initial studies conducted with college students (Lipnevich et al., 2014; Yang & Zhang, 2010) and high school (Lipnevich et al., 2022) showed promise with using exemplars after the initial draft has been submitted.

Our findings provided further support to the results reported in those studies and showed that annotated exemplars may enable comparisons and help learners to effectively self-assess and generate self-feedback that is conducive to improvement, with effects comparable to the ones from the teacher comments. Moreover, exemplars seem to work as well for middle-school students as they do for older students. The non-significant differences in the level of improvement in students' writing across conditions show great promise for practitioners. However, additional research is needed to test the consistency of these results in other samples.

Importantly, and this is something that all instructors will be able to appreciate, it took between eight and fifteen minutes to grade each essay, and it took about 30 min to annotate each exemplar, which can be re-used across different assessments and cohorts. Such dramatic time saving could help teachers to increase frequency of revised assignments without creating unbearable time and effort constraints. In other words, using annotated exemplars as feedback may improve student writing, may save teachers' time, and increase students' autonomy, as it helps learners to become less reliant on teacher support.

Interestingly, the combined condition, where students received both the annotated exemplars and teacher comments, did not outperform the other two groups. Prior studies that explored effects of rubrics and exemplars as feedback showed that participants in the combined condition (rubrics and exemplars) did not perform as well as the participants in the condition receiving only rubric (Lipnevich et al., 2014) or either rubric or exemplar alone (Lipnevich et al., 2022). In both cases, the researchers explained these findings through cognitive load and time limitations: It simply takes more time and effort to use both sets of instructional tools in the allotted time slot. In the current study, cognitive load could also be a viable explanation. It is possible that students did not take full advantage of both forms of feedback, likely focusing on the comments from the teacher, which they were more accustomed to receiving. The similarity of the effect sizes between teacher comments ($d = 1.21$) and the combined conditions ($d = 1.29$) offers some additional support to this explanation. That is, there was a virtually identical improvement between the teacher comments and the combined condition, with the effect for the exemplar group being somewhat higher ($d = 1.89$). Once again, we encourage the researchers to further explore the relative effectiveness of teacher comments, exemplars, and other instructional tools across a variety of tasks and settings while exploring additional explanatory factors (i.e., cognitive load).

6. Implications, limitations, and future directions

Our results present promising educational implications for both teachers and students. Teachers' choice to rely on annotated exemplars as a mechanism of feedback provision may offer a significant cut in the time necessary to deliver feedback – time that could potentially be spent on the development of creative educational activities, lesson planning, or research. Also, teachers could increase the number of writing and revision cycles expected from students, consequently increasing students' opportunities to exercise their writing skills. The use of alternative, more sustainable feedback practices can help students in the development of autonomy and self-regulatory abilities (Carless et al., 2011).

This study is not without limitations. We were not allowed to include a control group and only compared performance of students in three experimental conditions. Although we know that students struggle with revisions if no information is provided to them (Lipnevich & Smith, 2009), future studies exploring the effects of exemplars and comments may include a control group. Further, the sample included in this study presents a limitation for possible generalization. Because this study was conducted in a single private institution in Brazil where the participants were mainly white and from higher SES, results must be interpreted and generalized with caution. We encourage our colleagues to replicate this study in a wide variety of contexts. Also, our lack of control over the quantity and quality of students' engagement with different formats of feedback prevented us from controlling for the time-on-task variable. Future studies should control for the quality of students' interactions with their feedback in order to observe whether similar levels of active engagement with different formats could result in differential performance. In addition, more studies involving primary and middle school students could enrich the field, including longitudinal observations of the effects of the frequent use of exemplars as a form of feedback on students' writing tasks. Another interesting avenue for investigation could be in exploring how students cognitively engaged with teacher comments and annotated exemplars to generate self-feedback. Think aloud studies that capture students' processing of feedback during the revision process could shed light on those questions.

In conclusion, and acknowledging all the limitations of this study, we are excited about the possibilities presented by the current findings. Developing students' ability to self-assess and generate self-feedback at an earlier age through their interaction with annotated exemplars could lead to significant educational gains while also decreasing teachers' workload. We do not suggest that teacher comments should be abolished altogether in favor of standardized alternatives. However, the fact that in some cases we can have our students do the work on their own and request feedback if they think they need it or comment on their improved drafts later in the revision process, could be a welcome change to educators' teaching practices. We are certainly happy to use this practice in our own teaching.

Open Science

We report all data exclusions, all data exclusion criteria (if any), whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data

We confirm that there is sufficient information for an independent researcher to reproduce all the reported results, including the codebook and supplements can be retrieved from https://osf.io/tv8sf/?view_only=f8964c263911494dafa407597a579d03.

Open Materials

The information needed to reproduce all the reported methodology is openly accessible. The material can be retrieved from https://osf.io/tv8sf/?view_only=f8964c263911494dafa407597a579d03.

Preregistration of Studies and Analysis Plans

This study was not preregistered.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

We have no known conflict of interest to disclose.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2023.101262.

References

Andrade, & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 97*(1), 21–30. https://doi.org/10.1080/00220670309596625

Bell, A., Mladenovic, R., & Price, M. (2013). Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars. *Assessment and Evaluation in Higher Education, 38*(7), 769–788. https://doi.org/10.1080/02602938.2012.714738

Broadbent, J., Panadero, E., & Boud, D. (2018). Implementing summative assessment with a formative flavour: A case study in a large class. *Assessment and Evaluation in Higher Education, 43*(2), 307–322. https://doi.org/10.1080/02602938.2017.1343455

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407. https://doi.org/10.1080/03075071003642449

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.G. (2009). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences [Software]. Available from ⟨http://gpower.hhu.de/⟩.

George, D., & Mallery, P. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference 17.0 Update* (10th ed.). Boston: Pearson.

Graham, S. (2018). Instructional feedback in writing. In A. Lipnevich, & J. Smith (Eds.), *The Cambridge handbook of instructional feedback (Cambridge Handbooks in Psychology* (pp. 145–168). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316832134.009.

Graham, S., & MacArthur, C. (1988). Improving learning disabled students' skills at revising essays produced on a word processor: Self-instructional strategy training. *The Journal of Special Education, 22*(2), 133–152. https://doi.org/10.1177/002246698802200202

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2013). Teaching writing to middle school students: A national survey. *Reading & Writing, 27*(6), 1015–1042. https://doi.org/10.1007/s11145-013-9495-7

Graham, S., Harris, K. R., Adkins, M., & Camping, A. (2021). Do content revising goals change the revising behavior and story writing of fourth grade students at-risk for writing difficulties? *Reading & Writing, 34*(7), 1915–1941. https://doi.org/10.1007/s11145-021-10142-9

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Graham, S., & Perin, D. (2007). Writing next-effective strategies to improve writing of adolescents in middle and high schools. ⟨https://education.illinoisstate.edu/downloads/casei/5–15-WritingNext.pdf⟩.

Handley, K., & Williams, L. (2011). From copying to learning: Using exemplars to engage students with assessment criteria and feedback. *Assessment and Evaluation in Higher Education, 36*(1), 95–108. https://doi.org/10.1080/02602930903201669

Harris, K. R., & McKeown, D. (2022). Overcoming barriers and paradigm wars: Powerful evidence-based writing instruction. *Theory Into Practice*, 1–14. ⟨https://www.researchgate.net/publication/362295654_Overcoming_Barriers_and_Paradigm_Wars_Powerful_Evidence-Based_Writing_Instruction⟩.

Hendry, G. D., White, P., & Herbert, C. (2016). Providing exemplar-based "feedforward" before an assessment: The role of teacher explanation. *Active Learning in Higher Education, 17*(2), 99–109. https://doi.org/10.1177/1469787416637479

Holtz, J. W., & Daly, E. J. (2021). An evaluation of an instructional and motivational treatment package on writing revisions. *Contemporary School Psychology, 25*(2), 243–259. https://doi.org/10.1007/s40688-019-00247-y

Huang, & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition, 62*, Article 102675. https://doi.org/10.1016/j.compcom.2021.102675

King, D., McGugan, S., & Bunyan, N. (2008). Does it make a difference? Replacing text with audio feedback. *Practice and Evidence of the Scholarship of Teaching and Learning in Higher Education, 3*(2), 145–163.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4. https://doi.org/10.3389/fpsyg.2013.00863

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. https://doi.org/10.1007/s11251-013-9299-9

Lipnevich, A. A., & Panadero, E. (2021). A Review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.720195

Lipnevich, A. A., Panadero, E., & Calistro, T. (2022). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000434

Lipnevich, A. A., & Smith, J. K. (2009). The effects of feedback on student examination performance. *Journal of Experimental Psychology: Applied, 15*, 319–333.

Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn": Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*(4), 347–367. https://doi.org/10.1007/s11092-009-9082-2

Lipnevich, A. A., & Smith, J. K. (2022). Student–feedback interaction model: Revised. *Studies in Educational Evaluation, 75*, Article 101208. https://doi.org/10.1016/j.stueduc.2022.101208

MacArthur, C. A. (2018). Evaluation and revision. In S. Graham, C. MacArthur, & M. Hebert (Eds.), *Best practices in writing instruction* (pp. 287–308). New York: Guilford Press.

Mañez, I., Lipnevich, A., Lopera-Oquendo, C. & Cerdán, R. (in press). Don't be negative! Pre-service teachers' feedback on writing assignments does not look that bad. Teaching and Teacher Education.

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*. https://doi.org/10.1016/j.learninstruc.2022.101658

Nicol, D. (2020). The power of internal feedback: exploiting natural comparison processes. *Assessment and Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nicol, D., & McCallum, S. (2022). "Making internal feedback explicit: Exploiting the multiple comparisons that occur during peer review." doi:10.31234/osf.io/ksp2v.

Panadero, E., Lipnevich, A. A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In D. Boud, M. D. Henderson, R. Ajjawi, & E. Molloy (Eds.), *The impact of feedback in higher education: Improving assessment outcomes for learners*. Springer. https://doi.org/10.1080/02602938.2020.1823314.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). "Feedback: All that effort, but what is the Effect?". *Assessment & Evaluation in Higher Education, 35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Price, D., Smith, J. K., & Berg, D. A. (2017). Personalised feedback and annotated exemplars in the writing classroom" an experimental study in situ". *Assessment Matters, 11*, 122–144. ⟨https://www.nzcer.org.nz/node/60031⟩.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨https://www.R-project.org/⟩.

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. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209. https://doi.org/10.1080/0305498870130207

To, J., & Carless, D. (2016). Making productive use of exemplars: Peer discussion and teacher guidance for positive transfer of strategies. *Journal of Further and Higher Education, 40*(6), 746–764. ⟨https://web.edu.hku.hk/f/staff/412/2015_Making-productive-use-of-exemplars.pdf⟩.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174–1204. https://doi.org/10.1037/a0036620

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. https://doi.org/10.1080/00461520.2016.1207538

Yang, L., & Zhang, L. (2010). Exploring the role of reformulations and a model text in EFL students' writing performance. *Language Teaching Research: LTR, 14*(4), 464–484. https://doi.org/10.1177/1362168810375369