Contemporary Educational Psychology 62 (2020) 101897

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback

Yong Wu[*], Christian D. Schunn

Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, USA

ARTICLE INFO

Keywords:
Peer feedback
Feedback quality
Feedback frequency
Feedback features
Implementation

ABSTRACT

Prior research on the complex process of revision based upon peer feedback has focused on characteristics of each piece of feedback in isolation. Multipeer feedback allows for feedback to be repeated (or not), which could be a signal of feedback quality or be especially persuasive to peers. Separately, little research has examined how well peers select more impactful and accurate peer feedback in their revisions, whether repeated or not. We analyzed almost 2,000 peer comments received by 107 students in a secondary writing course in the US to determine whether feedback quality and feedback frequency predicted feedback implementation. Controlling for other feedback features and context factors, students were much more likely to implement feedback as both feedback quality and feedback frequency increased, surprisingly with no interaction (i.e., even low-quality comments were more likely to be implemented when repeated). However, low-quality comments often partially overlapped with high-quality comments, providing a potential explanation for the lack of an interaction. Finally, consideration of feedback frequency and feedback quality provides new insights into which feedback features are actually related to implementation. The results generally allay concerns about the blind-leading-the-blind in peer feedback as well as pushing for peer feedback arrangements that produce more overlapping comments.

1. Introduction

One instructional strategy that has drawn increasing attention is the use of peer feedback, defined as students at the same level giving each other formative feedback on classroom artifacts (e.g., essays or presentations). Sometimes also called peer assessment (with a focus on scoring) or peer review (including both feedback and scores), this strategy has been used across a wide range of student levels and content disciplines. In terms of the effects on the students receiving peer feedback, a number of benefits have been noted in comparison with teacher feedback: receiving more rapid feedback (Cho & MacArthur, 2010, Hovardas et al., 2014; Topping, 1998), receiving more detailed feedback (Cho & MacArthur, 2010), and receiving feedback that is more understandable to students (Cho et al., 2008; Hovardas et al., 2014). In spite of these advantages, peer feedback is regarded with suspicion and therefore avoided because of face validity concerns: peers are likely novices in the topic of instruction and therefore may not be able to provide quality feedback (Cheng et al., 2015; Kaufman & Schunn, 2011) or may not choose to put in the effort to provide quality feedback.

A number of process scaffolds have been proposed to address this concern: well-structured, student-friendly rubrics (Schunn et al., 2016; Wang, 2014), anonymity to allow for honest feedback (Lin, 2018), and grades for high quality feedback (Patchan et al., 2017). Another strategy involves the use of multipeer review (Cho & Schunn, 2018; Ge, 2019; Tseng & Tsai, 2007): if multiple peers give feedback, the odds that at least one reviewer detects key problems or provides strong advice will go up and the occasional poor advice might be ignored if only given by peers. Variability across peers in peer feedback (Hovardas et al., 2014) and cross-validation of feedback made across reviewers (Van Steendam et al., 2010) have been noted in the literature. However, the direct effects of receiving multiple comments on the same issue have received little attention (Gao et al., 2019).

Instead, prior research investigating factors that might influence the impact of peer feedback generally focused on one or two comment-level factors. For example, there has been considerable focus on the influence of features within how feedback is given (e.g., including praise, explanations) on feedback implementation (e.g., Leijen, 2017; Lu & Law, 2012; Patchan et al., 2016; Tseng & Tsai, 2007). However, feedback-based revision process is complicated and influenced by a variety of factors. Here we consider two potentially interacting factors that are central to multipeer feedback and that have received little attention in studies of how feedback influences revision processes: the quality of feedback and the frequency of feedback. Although it may seem obvious that these features are likely to influence peers, the size of the effects

[*] Corresponding author.
E-mail address: yongwu@pitt.edu (Y. Wu).

https://doi.org/10.1016/j.cedpsych.2020.101897

Available online 15 July 2020
0361-476X/ © 2020 Elsevier Inc. All rights reserved.

are unclear. If there are large effects, they may prove to be central factors in understanding when and why peer feedback is effective.

We present a study that systematically codes feedback quality, feedback frequency, and feedback implementation, and then examines the relationship of quality and frequency to implementation in the context of a secondary writing course. In this study, we also examine whether the inclusion of these factors changes understanding of a peer feedback research topic already examined in depth: the role of feedback features on feedback implementation. Thus, this research can provide new insights into the nature of multipeer feedback as well as suggestions for possible directions that might improve the effectiveness of peer feedback.

2. Theoretical background

2.1. Peer feedback quality

Different criteria have been used to evaluate peer feedback quality. Some have focused on the structural components of feedback. For example, Sluijsmans et al. (2002) developed a rating form of peer assessment quality including seven criteria: use of criteria, positive comments, negative comments, constructive comments, posed questions, naïve word use, and structure. Prins et al. (2006) elaborated on this framework to develop a feedback quality index, which includes nine items related to three broad variables: use of criteria (i.e., quality of both content and explanation), nature of feedback (i.e., remarks, questions posed, repertoire, and advice), and writing quality (i.e., structure, formulation, and style). This framework has been used to evaluate the effects of different scaffolds on peer feedback quality (e.g., Gielen & De Wever, 2015a). However, these studies did not directly validate the dimensions in terms of the actual effects of peer feedback (i.e., do peer comments with higher quality ratings have greater influence on peers?). Most importantly, these dimensions of feedback quality do not speak directly to the heart of student and teacher concerns: is the feedback accurate and focused on aspects of the artifact most needing to improve?

Other studies have taken the accuracy of feedback content into consideration when evaluating feedback quality. In a study by Van Steendam et al. (2010), students were required to identify errors in a given text including 10–20 flaws and suggest revisions. Peer feedback quality was evaluated in terms of “the correctness, exhaustiveness and explicitness of student comments” (Van Steendam et al., 2010: 321). This approach is not applicable to research conducted in a naturalistic and authentic learning environment because errors, especially high-level problems, are hard to exhaustively identify in student writing. Further, high-level issues can vary in importance to the document being evaluated, so detecting 80% of the problems may miss the most important 20%. Further, the feedback can vary in terms of the extent to which the advice fully addresses the problems detected. In other words, a suggestion for improvement can be accurate but not necessarily sufficient to substantially improve the document (e.g., giving one small suggestion for improving pervasive problems in essay organization).

Hovardas et al. (2014) used a hybrid approach, defining feedback quality according to both feedback structural components and feedback content accuracy. They also compared the quality of peer feedback and teacher feedback, and then investigated the influence of feedback quality on revisions. Interestingly, students did not respond to all accurate suggestions, and students also validated feedback by cross-checking across feedback sources, pointing to an important role of feedback frequency, not just feedback quality in uptake of peer feedback.

Since the structure of feedback (e.g., including positive and negative feedback, elaborated information, or suggestions; Gielen & De Wever, 2015a; Prins et al., 2006; Sluijsmans et al., 2002) might influence peers via a form of persuasion (Wu & Schunn, 2020), the structure of the feedback is not necessarily positive or negative in terms of influencing performance. For example, praise within feedback might persuade a peer to implement the feedback even if the feedback is not accurate. Therefore, we argue that is important to separate the structural aspects of peer feedback (how it is phrased; which we call feedback features) from the accuracy of problems detected/usefulness of suggested solutions (which we call feedback quality). In particular, we take an integrative functional approach to defining quality: will following the peer feedback produce a meaningful improvement in the document being evaluated? Such a definition is closest to the sense of accuracy that instructors and students will value as the measure of accuracy of feedback.

Such an approach (i.e., separately considering feedback quantity from feedback quality and feedback features) has not been previously taken in studying peer feedback. In the context of studying the uptake of peer feedback, it is unclear whether students will recognize which feedback is more important to follow and whether they will ignore bad advice, especially when it is persuasively structured (e.g., with strong supports such as explanations and detailed suggestions, or maybe repeated multiple times). Since students are generally using language that is understandable by peers, it is likely that they will prefer useful over useless advice, but the size of this preference is unclear.

2.2. Peer feedback frequency

Feedback frequency is an important and salient factor within multipeer review. Theoretically, it should be important. Agreement among reviewers on the central problems can persuade authors that the problem is real and the reviewers are trustworthy, whereas disagreement among the reviewers can have the opposite effect (Kaufman & Schunn, 2011). Also, when multiple comments (perhaps from multiple reviewers) attend to the same issue but with slightly different details (e.g., different amounts of information, different suggestions for revision), the author might be more likely to address the issue because more of their needs are met (e.g., in understanding the problem or in understanding how to address the problem).

Interestingly, few studies have studied feedback frequency directly. Hovardas et al. (2014) did not directly measure feedback frequency, but they found that students were more likely to make revisions if they found peer feedback and expert feedback they received agreed, which is one form of repeated feedback. Another related construct is amount of critical feedback received. Patchan et al. (2016) found that receiving more comments overall reduced implementation rates whereas receiving more critical comments increased implementation rates. These opposing effects could be related to capacity overload issues (i.e., too many total comments) or persuasion-to-act effects (i.e., there are significant problems, so some action is required).

Focusing more specifically on frequency of particular comments, Patchan et al. (2013) found that multipeer feedback was mostly complementary comments (i.e., each peer provides additional issues to address), but close to 10% of the comments were overlapping in content with similar recommendations. Leijen (2017) investigated whether feedback implementation was influenced by repetition of peer feedback, defined as “have other peers referred to the same/similar aspect in their feedback” (p. 42). It was found that repetition, as a binary variable, predicted feedback implementation significantly. Gao et al. (2019) studied the likelihood of repairing a problem in the document using more levels of repetition: none, 1, 2–4, and more than 4. Interestingly, more than four had the highest level of repair rates and receiving only one comment was equivalent to receiving no comment on the issue. However, only 21 students were involved in the study and high repetitions of comments were rare, so the statistical power was not strong.

Additional studies are required to more precisely investigate the effects of repetition with sufficient statistical power and take statistical confounds into account. In particular, it seems likely that feedback frequency will be confounded by feedback quality: important problems will receive more comments. Thus, to understand the effects of feedback frequency or feedback quality, the other factor must also be carefully controlled. It is likely that frequency will influence implementation rates, but the size of the effects is unclear, as well as it being unclear whether receiving more than just two comments will matter.

Another related question involves the interaction between frequency and quality. On the one hand, low quality feedback should not be implemented regardless of how often it is endorsed by reviewers. That is, normatively speaking, frequency should only matter in the case of higher quality feedback. On the other hand, at a process level, quality and frequency can be independent cues to revision and thus not interact (i.e., students make a revision when they notice it is useful or if they are persuaded by multiple peers making similar comments).

2.3. Peer feedback features

Although not a central focus of the current study, as the literature review above makes clear, it is important to take into account statistical confounds when studying feedback frequency and feedback quality, and feedback features are at the top of the list of important confounds. Further, since frequency and quality have been so rarely investigated in studies of feedback features, it may be that prior work has not accurately identified which feedback features are actually important for influencing uptake of peer feedback because frequency and quality have not been taken into account.

Feedback features (i.e., structural components of feedback), which are also called feedback content in some studies (e.g., Strijbos et al., 2010), have been regularly found to play a crucial role in the effects of peer feedback on student performance (Cho & MacArthur, 2010; Lin, 2018; Schunn & Wu, 2019). Feedback features have been divided into broad categories, such as evaluative vs. informational (Narciss, 2008), simple vs. elaborated (Narciss, 2008; Strijbos et al., 2010), or feedback with a varying structuring degree (Gielen & De Wever, 2015a). Prior research has found elaborated peer feedback is more beneficial than simple feedback (Gielen & De Wever, 2015a; Strijbos et al., 2010).

Other researchers divide peer feedback into more specific types: identification, explanation, suggestion, solution, and praise (Cho & MacArthur, 2011; Lu & Law, 2012; Nelson & Schunn, 2009; Tseng & Tsai, 2007). Such more specific feedback type frameworks, while more complex, can be more useful than the simpler frameworks because these more specific features are easier to describe to instructors and students (Gielen & De Wever, 2015a) and they appear to provide better predictions for what features influence peers uptake of the feedback they receive. For example, not all elaborations or structuring are equally helpful, as reviewed below. Rather, explanatory feedback in particular is found to be especially helpful in improving students' writing (Gielen et al., 2010), and thus reviewers are encouraged to specifically provide explanations of the problems they identify.

In terms of the influence of these features on students' writing performance, a number of different features have been studied. Some studies observed a positive correlation between including a suggestion and feedback implementation (Leijen, 2017; Nelson & Schunn, 2009; Patchan et al., 2016). But these studies did not distinguish general suggestions and more specific solutions. Wu and Schunn (2020) found that specific solutions were most helpful because they were especially likely to be understood.

A number of other studies have found a positive influence of explanatory feedback on receivers' implementation (Gielen et al., 2010; Patchan et al. 2016; Wu & Schunn, 2020), but this feature was sometimes not a significant predictor (e.g., Leijen, 2017) or even a significantly negative predictor (e.g., Nelson & Schunn, 2009; Tseng & Tsai, 2007) of writing performance, perhaps because learners as novice reviewers could not provide clear explanations.

With respect to praise feedback, some observed positive correlations between praise feedback and receivers' revised drafts (e.g., Lu & Law, 2012; Tseng & Tsai, 2007). However, Patchan et al. (2016) found that students were less likely to implement received feedback in the revised drafts when praise was included in a criticism comment. Note that these studies define praise feedback in different ways. Some (e.g., Patchan et al., 2016) distinguish mitigating praise (i.e., positive feedback included in a negative feedback to soften the criticism) and pure praise (i.e., pure positive feedback praising the work), while the others define it as positive feedback praising the work, which include both mitigating praise and pure praise (e.g., Lu & Law, 2012; Tseng & Tsai, 2007).

The mixed prior results for each of the feedback features might have resulted from different definitions of feedback features or from different learning outcomes, such as feedback implementation on the comment level (Nelson & Schunn, 2009; Patchan et al., 2016) or quality of revised drafts (Lu & Law, 2012; Tseng & Tsai, 2007). However, another possibility is that feedback quality and feedback frequency were not properly controlled. The sometimes negative effects of some of the features suggest that feedback quality may be important.

2.4. Additional variables

Other variables might also influence feedback implementation and thus need to be statistically controlled in examining relationships of frequency and quality with implementation. Here we discuss prior work on several such variables that are relevant to the current study. When drawing on data from multiple schools, it is important to control for the potential effects of school because environment-related factors, such as prior educational experiences of students, academic performance of peers, teaching experience of teachers, norms and expectations from teachers, are likely to influence how consistently problems are noticed and how commonly revisions are made (Hughes, 2012), and thus producing a large potential confound in statistical models predicting implementation (Wu & Schunn, 2020).

Similarly the quality of the document receiving comments simultaneously influences the number and diversity of comments received and the likelihood of implementing suggested revisions. When students receive a strong evaluation on their first draft (Wu & Schunn, 2020) or few comments that require document revisions (Patchan et al., 2016; Wu & Schunn, 2020), they are less likely to perceive a need to make any revisions.

Relatedly, there is one superficial indicator of comment quality that can influence implementation: spelling errors in the comments. Readers may holistically judge others as overall incompetent when they encounter spelling errors. In peer feedback, students already have a concern about some of their peers not having sufficient competence to provide accurate feedback (Kaufman & Schunn, 2011; Yang et al., 2006). Indeed, in the current study, some students specifically complained about the negative effects of spelling errors in peer feedback (e.g., "*The reviewer has many spelling errors throughout their review, which sometimes makes it confusing to understand what exactly they're trying to convey to me*".).

Additional potential correlates of comment quality occur at the student level. For example, student gender can be a confound to control in analyses of peer feedback on writing. Girls often outperform boys in writing performance across all grades (Reilly et al., 2019), which likely also relates to comment quality. Stylistically, girls sometimes tend to include different comment features (Prinsen et al., 2009) and overall give more comments (Hamer et al., 2015), which would be related to comment frequency. Girls might also tend to respond to peer feedback in different ways than do boys (Prinsen et al., 2009). Similarly, older students might perform differently in peer feedback because older students tend to be stronger writers and may be more experienced or mature in responding to feedback.

2.5. The present study

In the present study, we integrate and extend existing research in peer feedback by investigating the influence of three major comment-level variables (feedback quality, frequency, and features) on feedback implementation in the naturalistic setting of a secondary school writing course. Although an extensive body of feedback research has examined

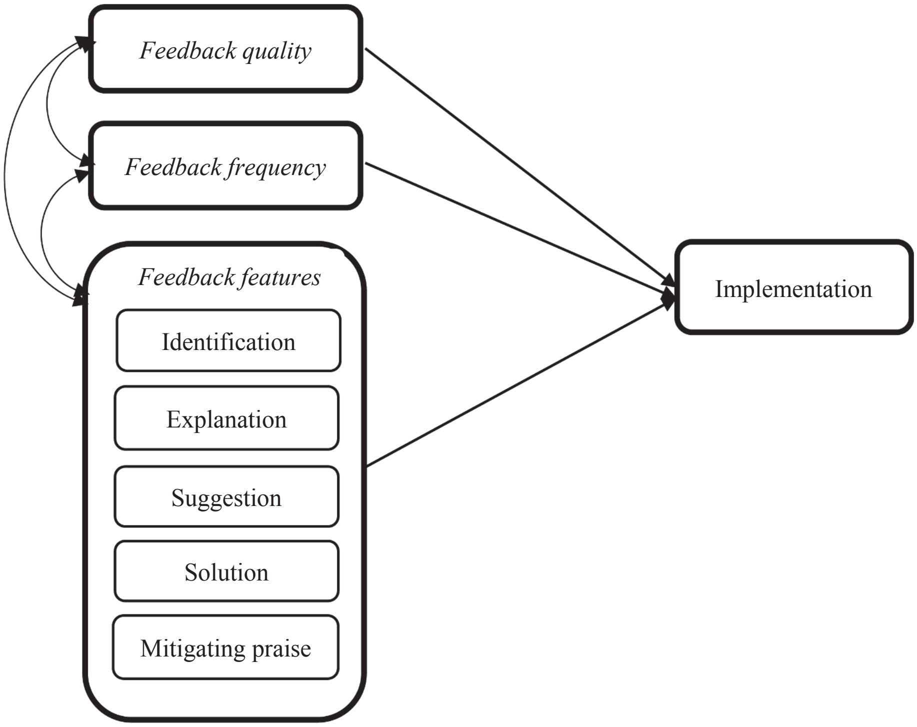

Fig. 1. Proposed feedback-to-implementation model. Straight lines with single-headed arrows represent causal relationships whereas curved lines with double-headed arrows represent correlated variables.

the relationships of feedback features or feedback quality and revised draft scores, little research has examined how feedback quality and frequency influence its implementation. By investigating all three variables, along with other important controls, the actual effects of each variable may be better revealed in a naturalistic setting. Second, studies on peer feedback at the secondary school levels are lacking (Hovardas et al., 2014; Schunn et al., 2016). As a result, very little is known about what secondary school students can do in a multipeer feedback context (e.g., their feedback quality, frequency, and features).

The conceptual model of the hypothesized relationships of feedback quality, frequency, and features to feedback implementation is presented in Fig. 1. All three (quality, frequency, and features) are conceptualized as potential causes of decisions to implement received peer feedback in a revision. Feedback frequency is conceptualized as being correlated with feedback features and feedback quality, whereas feedback quality and frequency may influence the relationships of feedback features to implementation.

Two major research questions are addressed in the present study:

RQ1: *What are the relationships of feedback frequency and feedback quality (overall and interactively) to feedback implementation?*

RQ2: *Does controlling for feedback quality and feedback frequency change the observed relationships between feedback features and feedback implementation?*

3. Method

3.1. Overview

To address the research questions, the approach was to 1) code the quality, frequency, and features of each received peer comment; 2) code implementation of each comment in the document revision; 3) analyze the relationships of comment quality, frequency, and features to implementation. Further a sequential quantitative/qualitative approach (Ivankova et al., 2006) was used to address each research question, first uncovering quantitative patterns and then further exploring the nature/possible alternative explanations of the observed quantitative patterns.

3.2. Participants

Participants were students enrolled in Advanced Placement (AP) Language and Composition. AP courses are high school courses that are meant to be equivalent to common introductory university courses. This course is the most commonly taken AP course, and it has the aim of helping students develop skills related to analytic and persuasive writing (see College Board, 2018). However, this course's mean overall score on the 1 to 5 scale for all AP courses has been consistently below a 3 (the lowest threshold universities ever use for accepting it as a replacement for their own courses), one of the lowest mean scores among all the AP courses. This low performance level as well as the very large number of students enrolled in the course creates a need for more research on how to better help these students better prepare for college-level writing.

One hundred and seven secondary school students (mean age = 17.4 years, *SD* = 0.5; 63 female) participated in the study. Of the participants, 51% were Caucasian, 14% were Asian, 14% were Hispanic/Latino or African American, and 21% did not report their ethnicities. Students were enrolled at six secondary schools from across the United States, selected to represent a range of school and student demographics (e.g., urban and rural, public and private, high and low socio-economic status). 50% of the participants came from Title I schools, which serve many low-income families and receive additional federal financial support to help students of low-income families (DOE, 2018).

Teachers teaching a wide range of schools were recruited through email for participation in the study. Volunteer teachers were selected if they had previously taught the AP course, had at least two sections of students, and were willing to implement the core requirements of the study: implement a shared writing assignment with peer review using a common online tool and common rubrics at a common time during the school year. In each school, a single teacher taught the multiple course sections participating in the study.

3.3. Materials

Peer assessment tool. This study is part of a larger project in which an online peer assessment program, *Peerceptiv*, was used to support writing instruction and assessment (Cho & Schunn, 2007; Schunn et al., 2016). The program is being widely used at the secondary and tertiary levels throughout the world. It involves a web-based platform on which students can upload their assignments and perform peer review in groups anonymously according to rubrics specified by the teacher. The program allows teachers to assign writing tasks and monitor student performance. Additional functions are included to improve feedback quality, such as well-designed rubrics, accountability measures, and suggestions for including helpful information in feedback (Patchan et al., 2017).

Writing task. The writing task submitted to peer review involved an evidence-based argument, and was taken from prior end-of-course AP exams. To write the essay, students were required to read a one-page passage, and then write a paper analyzing the passage, discussing the rhetorical strategies the passage's author used while providing support evidence for use of these strategies from the source passage. Students were given one week to complete the essay.

3.4. Measures

Four undergraduate research assistants coded all the feedback features and feedback topics (which was then used to compute feedback frequency). Two writing experts coded feedback implementation and feedback quality. Everything was independently and exhaustively double-coded; conflicts were discussed and resolved to raise the effective reliability of the resulting data used in analyses. The measures are summarized in Table 1.

Feedback segmentation and implementability. To more precisely examine the effects of feedback, all the feedback comments were first segmented into idea units; one reviewer might give several revision ideas to a given essay on a given dimension. A separate feedback comment has a unified intended purpose and worked toward a given function. For example, the following comment was divided into the two idea units: *"Your explanation could use more details, although the details you did provide make sense and support your claim. Just work on adding more to your analysis and going deeper into your writing. // The word choice is good but it could have been better"*. The sentences before the double slash indicated a high-level problem, and the last sentence after the double slash identified a low-level problem.

Next, comments were coded for whether they were implementable or non-implementable (*Kappa* = 0.92). Implementable comments could trigger revisions. Non-implementable comments could not be used in revisions: comments that consisted of only positive comments or only a summary of what the writer did. Praise and summary comments were removed from further data analyses, since there was nothing to implement in either of those cases. These two steps yielded a total of 2,057 implementable feedback comments.

Feedback level. Feedback comments were coded for their focus on high-level vs. low-level aspects of writing (*Kappa* = 0.91). Although not the focus of the research questions, comments about low-level language issues were expected to have different rates of features (e.g., more solutions) as well as be more likely to be implemented because those changes are easier to make. Such differences are statistical confounds for the core research questions that need to be addressed in the statistical models. Note that although there were separate reviewing prompts for these aspects of the essay, reviewers did not always stick to only the requested information within a given comment, and thus feedback level needed to be coded. A comment was coded as High-level if the comment concerned the thesis, the argument in the essay being analyzed, the rhetorical strategies, evidence for claims, explanations about the evidence, and organization. Low-level comments were those that focused on control of language and conventions.

Feedback implementation. To code whether a comment was implemented, MS Word's Compare Document tool was used to locate the changes between students' first and second submitted drafts. Coders compared each piece of feedback to the changes in the document to determine whether changes were made that aligned to the peer feedback received (*Kappa* = 0.74 for Implemented or Not Implemented). A comment was labeled as Implemented if a change could be interpreted as attempting to respond to a comment. If the comment did not appear to result in any revision, it was "Not Implemented". Not all comments could be coded for implementation: about 6% of the comments (*N* = 41 high-level and *N* = 88 low-level comments) were vague so that it was difficult to determine whether they were implemented or not (e.g., "*The style is very simplistic*."). These vague comments were removed from further coding and analysis.

Feedback features. Feedback features involve feedback components that can be included in a comment. Based on prior research arguing for the importance of particular features in peer feedback (e.g., Leijen, 2017; Lu & Law, 2012; Tseng & Tsai, 2007), each implementable comment was coded for whether it contained the following features: Mitigating Praise (*Kappa* = 0.85), Identification (*Kappa* = 0.81), Explanation (*Kappa* =0.80), Solution (*Kappa* = 0.76), and Suggestion (*Kappa* = 0.79) (see Table 1 for definitions and Table 2 for examples).

Feedback quality. Feedback quality was coded by two independent raters based on the degree to which addressing a given comment (if addressed as the reviewer suggests) would likely improve essay quality in measurable or non-measurable ways on the essay rubrics (*Kappa* = 0.69, see Table 3 for definitions and examples). Feedback quality was coded as (1) High-quality (*N* = 588); (2) Medium-quality (*N* = 810); (3) Low-quality (*N* = 530). The Low-quality feedback category was created by combing three categories that proved to each be relatively low frequency: No Effect (*N* = 113), Mixed Effect (*N* = 365), and

Table 1
Definitions of the outcomes, control/contextual, and feedback variables examined in the study.

Variable	Type	Level	Definition
Outcome			
Implementation	Binary	Comment	Whether a comment was implemented
Feedback quality	Binary	Comment	Whether a comment improved essay quality
Feedback frequency	Continuous	Comment	Frequency of each feedback topic an author received
Feedback features			
Mitigating praise	Binary	Comment	Whether an implementable comment included mitigating praise
Identification	Binary	Comment	Whether a problem was identified explicitly
Explanation	Binary	Comment	Whether an implementable comment included an explanation
Solution	Binary	Comment	Whether an implementable comment included a specific solution for revision
Suggestion	Binary	Comment	Whether an implementable comment included a general suggestion for revision
Control and contextual			
Number of implementable comments	Continuous	Author	Total number of implementable comments
Number of spelling errors	Continuous	Comment	Number of spelling errors in a comment
Author's task 1 score	Continuous	Author	Ratings of task one first draft quality
Gender	Binary	Author	Whether a student is female
Age	Continuous	Author	The age of an author
School title	Binary	Author	Whether students came from Title I school

Table 2
The five coded feedback features with examples.

Category	Examples
Mitigating praise	*You do an amazing job of explaining your real-world examples. In your third body paragraph, however, you add too much biased information for your point.*
Identification	*Your essay has no pattern; it was hard for me to read.*
Explanation	*In paragraph three the evidence was inaccurate; it was not a personal experience Louv was describing in the text, it was something that happened to his friend.*
Solution	*In your thesis statement, you mention Louv's bias but you fail to state what his bias is. Maybe you could add to the end: "…and rhetorical questions to emphasize his argument against the separation of nature and society."*
Suggestion	*Adding an extra rhetorical device will give you more to talk about and just strengthen everything.*

Table 3
Coding scheme with examples for feedback quality codes.

Category	Description	Examples
High-quality	A comment could improve essay quality measurably on rubrics.	*The introduction does not clearly state which rhetorical devices exactly will be analyzed, and therefore it's hard to tell which rhetorical devices you are analyzing.*
Low-quality	A comment could improve essay quality but not measurably on rubrics.	*All of your paragraphs start the same way, "Louv…". Maybe you should consider changing it up, so it will be less commonly used.*
No effect	A comment could not influence essay quality.	*The thesis statement is well-written and provides a solid statement to go off of. One thing I would suggest for that is to either add a word like "describe" before saying "why that is bad" or to get rid of it and allow your body paragraphs to show what side Louv is on.*
Incorrect	A comment would harm essay quality.	*Was not a thesis.*
Mixed effect	A comment could improve essay quality in part, and hurt essay quality in part.	*Even though there was evidence taken from the text, there still could have been more quotes put into the composition.*

Incorrect (N = 52). The Mixed Effect case occurs when a comment implies multiple changes should be made to address a given problem, and the separate changes would be placed in different categories (e.g., High and Incorrect or Low and No Effect).

Feedback frequency. To measure feedback frequency, a framework was needed for determining when two comments were the same. Such a determination is far from trivial since there are many ways of giving the same advice in open text comments. Further, one student may receive over 50 peer comments, and coding repetition by comparing every comment to every other comment is a very slow process. The approach taken was to code comment topics on a detailed level (e.g., Missing evidence) and then count topic frequency. Early cross-validation work based on a subset of the data showed that this approach closely approximated frequency estimates obtained using more time-consuming methods.

A detailed, assignment-specific feedback topic coding scheme was developed inductively to cover all the main topics found in the peer feedback. First, 200 randomly-selected high-level comments were coded for the specific problem indicated in each comment. For example, "*Adding an extra rhetorical device will give you more to talk about and just strengthen everything*" was coded as "Insufficient rhetorical devices". The first round of coding generated an initial feedback topic coding scheme. Not surprisingly, the topics were strongly aligned with each of the review rubric dimensions. Second, another 300 high-level comments were coded based on the initial coding scheme. It was found that some topics could be combined (e.g., General thesis & Vague thesis), and some new topics needed to be added. The coding categories were augmented and refined until almost all the feedback topics were covered (i.e., relatively few comments were placed in an "other" category; see Appendix B for examples). Low-level feedback topics required additional divisions because the specific problems are basically separate issues to be addressed (e.g., misuse of pronouns and incorrect sentence structures are different problems to address even though they are both in grammar topic). Third, the full set of comments were double-coded by four trained coders (*Kappa* = 0.75 for topic coding). Finally, when all the feedback-topic coding was done, the frequency of comments for each topic was calculated for each student as an author. For analysis purposes, three frequency categories were used: once, twice, and three+.

Additional variables. To more effectively reveal the unique relationships between feedback quality, frequency, features, and implementation, six potential confounds based on prior research on peer assessment and writing were included as control variables in the statistic models (see Table 1). Number of implementable comments was calculated by aggregating all the implementable comments each author received. Spelling errors in comments were initially identified using the Microsoft Word spell checker and then verified by one coder. Author's Task 1 scores were obtained from grading done by two writing experts with years of teaching experience who were trained to use the peer rubrics (see Appendix A), and they did so with substantial interrater reliability (*Kappa* = 0.72).

Procedure. All teachers participating in the study agreed to use a shared writing assignment for the peer review task, shared peer review rubrics, and a shared student training process. Prior to the commencement of the study, the teachers were trained on how to use the peer assessment system and rubrics. The particular writing task was drawn from prior end-of-course AP exams and thus was generally familiar to the teachers. Teachers were also provided with shared protocols for training students on how to conduct peer review based on rubrics. The peer review rubrics were based on the rubrics used in the end-of-year AP exams, but were adapted to be more student-friendly and indeed produced peer ratings that are highly correlated with expert ratings (Schunn et al., 2016). The teachers had agreed to use the peer review system in their AP writing classes at a similar time of year.

Peer review training of students was conducted in class (see Appendix A). Two sample essays were analyzed and discussed collectively in the lesson to demonstrate what makes feedback helpful and the strategies students could use to comment on students' essays. In the first phase, students were presented a sample essay and example comments for the essay. Then students discussed as a whole class which example comments were helpful or not helpful, and what made the comments helpful. In the second phase, students were required to provide feedback on the second sample essay with a partner according to the rubrics. Then the whole class discussed the written comments and gradings they produced. In addition, they held a discussion about the rubrics and received calibration feedback to better understand what is expected from them.

The peer review process involved four phases. First, students wrote the first draft of the essay and submitted it to the system before the deadline set by the instructor. Then, the system randomly and anonymously distributed each essay to four peers within the school. Students as reviewers provided numeric scores and written feedback to the essays based on eight seven-point rating rubrics (see Appendix A) within one week. At least one comment had to be provided on each evaluation

Table 4
Percentage of comments (and raw number of comments) that are about high- and low-level writing issues at each combination of comment quality and comment frequency.

Quality	Frequency			Overall
	Once	Twice	Three +	
High Level				
High	11% (181)	10% (158)	14% (214)	35% (553)
Medium	14% (218)	10% (154)	12% (190)	36% (562)
Low	13% (211)	6% (92)	10% (164)	29% (467)
Overall	38% (610)	26% (404)	36% (568)	100% (1,582)
Low Level				
High	6% (22)	2% (8)	2% (5)	10% (35)
Medium	41% (143)	22% (77)	8% (28)	71% (248)
Low	12% (40)	5% (18)	1% (5)	18% (63)
Overall	59% (205)	29% (103)	11% (38)	99% (346)

Note: "Three + " means that the topic was mentioned three or more than three times. Percentages may not total to 100 due to rounding.

dimension. Finally, students revised the first draft according to the feedback received to produce a final version of the essay and submitted it to the system.

3.5. Data analysis

The 107 students received 1,928 comments, of which 1,582 were high-level and 346 were low-level. For high-level feedback, quality and frequency were not strongly correlated and all quality and frequency categories were observed with relatively similar frequency (see Table 4); thus, there was a good opportunity to test the separate and interactive effects of quality and frequency on implementation. By contrast, low-level feedback was dominated by medium-quality comments and comments made by only one peer, limiting the ability to analyze the effects of quality or frequency and especially their interaction (see Table 4). Further, since most of the rubric dimensions focused on high-level comments, the majority of comments were high-level comments. Thus, for lack of statistical power, low-level comments were excluded in the remaining data analyses.

Since there is a nested structure to the comment data (i.e., comments are nested within authors), a two-level hierarchical modeling approach was used. The examined variables were divided into comment-level variables (i.e., Level 1) and author-level variables (i.e., Level 2). Because of the small sample size per school (i.e., 107 students sampled from 6 schools), school could not be treated as a third nesting level. Instead, school effects were modeled via dummy codes for each school and included as Level 2 covariates. Comment-level and author-level binary variables (i.e., five feedback features, quality, school, gender) were added as uncentered variables. Comment-level and author-level continuous variables (i.e., feedback frequency, number of spelling errors, authors' task 1 score, number of implementable comments, age) were grand-mean centered. We used an Adaptive Gaussian Quadrature method of estimation to conduct the two-level logistic regressions in Stata version 15[1].

Logistic regression was used because of the binary outcome variable, thus the results of the models were presented as odds ratios (OR). An OR greater/below than 1 means a positive/negative relationship between the predictor and the outcome (when holding all other variables constant). An OR of 1 indicates no difference between the groups when other variables are included. Note that the proportion of Level 2 variance to the total variance is not as informative for multilevel logistic models as in multilevel linear models because Level 1 variance is heteroscedastic (Raudenbush & Bryk, 2002) and the Level 1 and Level 2 variances "are not directly comparable" (Merlo et al., 2006: 291).

A fit of each partial (i.e., control variables and some predictors) and full model (i.e., control variables and all of the predictors) was compared to a baseline model (i.e., only control variables) based on the AIC (Akaike information criterion) (Patchan et al., 2016). All models successfully converged. The statistical model examining the effects of feedback quality and frequency on feedback implementation is as follows, where we have comments at Level 1 for 1…*ij* cases nested within Level 2 by 1…*j* authors.

Level 1 model. Comment-level predictors are: feedback quality, frequency, features, and number of spelling errors. The Level 1 model is given by:

$$\eta_{ij} = \beta_{0j} + \sum\nolimits_{p=1}^{P} \beta_{pj} x_{pij}$$

where η_{ij} are the log-odds estimates of the likelihood of implementing a comment *i* from an author *j*, β_{0j} is the average log-odds estimates of the likelihood of implementing a comment within the author *j*, β_{pj} ($p = 1, 2\ldots, P$) are comment-level coefficients for each predictor *p* from an author, and x_{pij} are comment-level predictors for each comment *i* from an author *j*.

Level 2 model. Author-level predictors are: authors' task 1 score, number of implementable comments, age, gender, and school. In particular, the Level 2 model is:

$$\beta_{qj} = \gamma_{q0} + \sum\nolimits_{s=1}^{s_q} \gamma_{qs} w_{sj} + u_{qj}$$

where γ_{q0} is the intercept, the expected log-odds estimates of the likelihood of implementation β_{qj} when all predictors are set to zero, γ_{qs} ($q = 0 \ldots, Q$, $s = 1, 2 \ldots, S$) are the fixed effects of author-level predictors w_{sj} ($s = 1, 2 \ldots, S$), u_{qj} ($q = 0 \ldots, Q$) are the random effects of β_{qj}.

4. Results

RQ1: What are the relationships of feedback frequency and feedback quality (overall and interactively) to feedback implementation?

To answer the first research question, the relationship of feedback frequency and quality to implementation was explored descriptively through graphing. Mean implementation rates as a function of feedback frequency and feedback quality are shown in Fig. 2. At all levels of quality, implementation rates were higher when there were more comments on that topic. Similarly, at all levels of feedback frequency, implementation rates were highest for high-quality comments and lowest for low-quality comments. In other words, there was no qualitative interaction between the two effects. Looking at Fig. 2, there was at most a small quantitative interaction, with medium quality comments showing the highest sensitivity to frequency.

Two-level logistic regressions were run to test the statistical significance of the relationship of frequency and quality to implementation, and insure they were robust relationships when including various controls (see Table 6). The frequency categories were treated as linear predictors given that frequency is inherently an interval scale and the strongly linear relationships shown in Fig. 2. To improve statistical power and further simplify these complex analyses, medium-quality feedback and high-quality feedback were collapsed into a Good feedback category because they showed similar patterns in terms of their effects on feedback implementation and both categories involved positive improvements to the document; similar results were obtained with more complex models that included all three quality levels.

Correlations among comment-level variables and implementation were examined to reveal potential confounds and multicollinearity issues among the variables (see Table 5). Identification and explanation, in addition to quality and frequency, were correlated with implementation. Suggestion was correlated with quality. In Table 6,

[1] We also used the software package HLM 8 to perform the two-level logistic regressions and got the same results. We present the results from Stata version 15, primarily because it provides more measures for comparing models (e.g., AIC). In addition, both Adaptive Gaussian Quadrature and the Laplace estimation methods were tried, and the two methods yielded similar results.

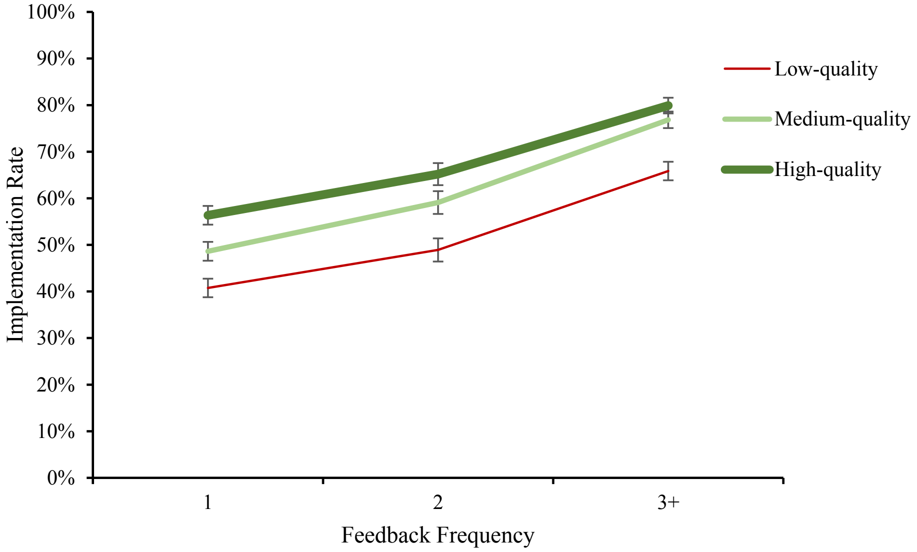

Fig. 2. Mean feedback implementation rates (with SE bars) as function of feedback frequency and feedback quality, for high-level comments ($N = 1{,}582$).

Table 5
Feedback quality, frequency, features and implementation: correlations for high-level comments.

	Variable	*M*	*SD*	1	2	3	4	5	6	7
1	Feedback quality	0.71	0.45							
2	Feedback frequency	1.97	0.86	0.07**						
3	Mitigating praise	0.47	0.50	0.02	−0.03					
4	Identification	0.73	0.45	−0.06*	−0.04	−0.22**				
5	Explanation	0.36	0.48	0.04	−0.02	−0.16**	0.43**			
6	Solution	0.13	0.34	−0.00	−0.11**	−0.01	0.00	−0.01		
7	Suggestion	0.72	0.45	0.12**	0.06*	0.05	−0.32**	−0.18**	−0.11**	
8	Implementation	0.61	0.49	0.13**	0.23**	−0.00	0.08**	0.10**	0.02	0.03

Note. $N = 1{,}582$.
*$p < .05$, **$p < .01$.

Model 1a is the baseline model including only control variables. Model 2a tests the two main effects of quality and frequency on implementation while including controls. Model 3a includes an interaction term to formally test whether the frequency and quality effects were purely additive. Model 4 is the full model adding controls for features within the feedback comment itself. The full model (Model 4 in Table 6) provided a better fit to the data than did the baseline model (Model 1a in Table 6): $\chi^2(7) = 109.8$, $p < 0.001$.

The effects of both feedback quality and frequency were consistent and robust (Models 2a–4 in Table 6). After adjusting for the control variables and feedback features, the association of feedback quality and feedback frequency with implementation remained. Students were more likely to implement a comment when it was good quality or repeated. The interaction between quality and frequency was small and not statistically significant, mirroring what appeared to be the case in Fig. 2: the effects of quality and frequency were independent effects. Further, random effects of Level 1 slopes were tested based on the full model[2]. Only the effects of feedback frequency on implementation varied across authors. Gender was found to be the only significant variable that influenced the relationship between feedback frequency and implementation ($B = 0.48$, $SE = 0.24$, $Exp(B) = 1.61$, $p = 0.04$). Girls seemed to be more likely to implement feedback than were boys when feedback was mentioned more frequently.

A follow-up qualitative analysis was conducted to explore the most surprising case from the non-interaction of frequency and quality: why did students implement low-quality advice when it was mentioned more than once ($N = 143$)? The first reason might be that the low-quality comments partially overlapped good feedback including both high-quality and medium-quality comments. That is, multiple comments across reviewers might focus on the same problem or the same aspects of the problem (hence being more frequent), but some of the reviewers might have given good advice while others gave bad advice for the same issue. Such overlap was formally coded (*Kappa* = 0.93 for overlapping or not) and found to be a frequent occurrence ($N = 104$). That is students would often made revisions that addressed low-quality comments that were repeated and overlapped good quality comments. For example, the two feedback comments were focused on the problem of needing to use better transitions at the beginning of paragraph 2 (see Example A in Appendix C). Comment 1 was rated as mixed perhaps because the solution was not appropriate. Comment 2 identified the general problem about transitions for each body paragraph, and then provided a specific solution for paragraph 2. This comment was rated as medium-quality. The author responded to these two comments by incorporating the suggested solution, thereby resulting in a low-quality comment (and a medium-quality comment) being coded as implemented.

An alternative explanation is that students might also implement low-quality comments because the comments were repeated (i.e., purely from persuasion). Much more rarely did students actually implement when they were repeated and only contained low-quality advice in all the repetitions: $N = 2$ incorrect and $N = 37$ mixed effects. Fifteen such incorrect or mixed effect implementations were of medium frequency (i.e., mentioned twice), and twenty-four of them were of high frequency (i.e., mentioned more than twice). For example, two

[2] The tests of random effects of Level 1 slopes were not presented in Tables 6 and 7 because they were not primary research questions.

Table 6
Two-level logistic regression analysis of quality, frequency, and implementation.

Predictors	Control variables (Model 1a)		Feedback quality and frequency (Model 2a)		Feedback quality, frequency, and the interaction term (Model 3a)		Feedback quality, frequency, and features (Model 4)	
	B (SE)	*Odds ratio*	*B (SE)*	*Odds ratio*	*B (SE)*	*Odds ratio*	*B (SE)*	*Odds ratio*
Mean outcome in logits (γ_{00})	−0.27 (0.62)	0.77	−0.71 (0.63)	0.49	−0.71 (0.63)	0.49	−1.29 (0.67)	0.28
Feedback quality	–	–	0.51 (0.15)	1.66***	0.51 (0.15)	1.67***	0.52 (0.15)	1.68***
Feedback frequency	–	–	0.74 (0.09)	2.09***	0.68 (0.14)	1.98***	0.76 (0.09)	2.14***
Quality*frequency	–	–	–	–	0.08 (0.17)	1.08	–	–
Feedback features								
Mitigating praise	–	–	–	–	–	–	0.15 (0.15)	1.16
Identification	–	–	–	–	–	–	0.37 (0.18)	1.45*
Explanation	–	–	–	–	–	–	0.41 (0.17)	1.51*
Solution	–	–	–	–	–	–	0.29 (0.21)	1.33
Suggestion	–	–	–	–	–	–	0.15 (0.18)	1.17
Control variables								
Total implementable comments	0.06 (0.04)	1.06	0.05 (0.04)	1.05	0.05 (0.04)	1.05	0.05 (0.04)	1.05
Number of spelling errors	−0.10 (0.06)	0.91	−0.12 (0.06)	0.89	−0.12 (0.06)	0.89	−0.15 (0.07)	0.86*
Author's first draft quality	−0.29 (0.24)	0.75	−0.26 (0.24)	0.77	−0.26 (0.24)	0.77	−0.23 (0.24)	0.79
Gender (F = 1)	0.61 (0.41)	1.83	0.69 (0.41)	2.00	0.69 (0.41)	2.00	0.65 (0.42)	1.92
Age	0.03 (0.39)	1.03	−0.02 (0.39)	0.98	−0.02 (0.39)	0.98	−0.04 (0.39)	0.96
Schools 2–6 (reference: School 1)	ns.		ns.		ns.		ns.	
Model fit statistics								
Log Likelihood	−846.7		−800.7		−800.6		−791.8	
AIC	1717.5		1629.4		1631.2		1621.7	

Note. N = 1,582. "-" means that the variables were not included in the model.
*p < .05, **p < .01, ***p < .001, *ns.*: not significant.

comments focused on the problem of the author failing to include a thesis (see Example B in Appendix C). Both of these two comments were incorrect. These two low-quality comments did not seem to overlap with any good comments.

RQ2: Does controlling for feedback quality and feedback frequency change the observed relationships between feedback features and feedback implementation?

Three two-level logistic regression models (see Table 7) were conducted to answer the second research question. Model 1b tested the relationship of feedback features to implementation without consideration of frequency or quality (but including other control variables). Models 2b and 3b tested the relationship of feedback features to implementation when controlling for feedback quality alone or frequency alone. Model 4 was the full model including both feedback quality and frequency; note Model 4 in Tables 6 and 7 are identical.

When feedback quality and feedback frequency were not included in the regression (Model 1b), explanation was the only significant feedback feature predictor of implementation. A comment including explanatory information was 1.52 times as likely to be implemented than a comment without any explanation. Adding feedback frequency in the model (Model 3b) did change the estimated relationships of feedback features with implementation (i.e., only explanation was a significant predictor). When feedback quality alone (Model 2b) or both feedback quality and frequency (Model 4) were included, explanation continued to predict implementation, with very little change in size. However, identification became a significant predictor. In other words, the inclusion of feedback quality changed the apparent relationship of identification to implementation.

Feedback quality was negatively correlated with identification (see Table 5), providing a possible explanation for why identification effects were masked when quality was not included in the model. Unpacking the more specific quality categories, this negative correlation primarily involved appropriate changes that produced non-measurable improvements (69% of identified comments) rather than incorrect, mixed effect, or no effect (3%, 23%, and 5% of identified comments, respectively). To better understand the relationships of identification to implementation, the relationship of identification across different quality levels, an interaction graph was created (see Fig. 3).[3] At both levels of quality (i.e., low-quality versus good-quality feedback), implementation rates were higher when identification was included in the comments. Thus, it appears that there is small but consistent role of problem identification in implementation that is hidden when not considering the large role of feedback quality.

5. Discussion

Fig. 4 summarizes the overall findings regarding peer feedback quality, frequency, and features with implementation.

RQ 1: What are the relationships of feedback frequency and feedback quality (overall and interactively) to feedback implementation?

Both feedback quality and frequency predicted feedback implementation, and feedback quality did not interact with frequency. It should come as no surprise that students are more likely to incorporate comments in revisions when the comments are repeatedly mentioned. However, the size of the effect is noteworthy: students were 2.14 times as likely to implement repeated comments. These research findings are consistent with Hovardas et al. (2014) who found that secondary school students take the decision-making strategy of triangulating peer and expert feedback received. They are more likely to make changes if there is overlap between peer and expert feedback.

Similarly, it should not be surprising that students are able to show some sensitivity to feedback quality, but the relatively large effect size of feedback quality (along with the relatively low base-rates of low-quality feedback) should be encouraging to teachers who are nervous about using peer feedback. Because these two effects are so large and that these two factors are correlated with other feedback factors, it will likely be important for future research on multipeer feedback to carefully take into account both of these factors, either as a potential confound (e.g., in the case of comment frequency) or as a potential

[3] We tested two interaction terms in the present study, but failed to find any significant effects of these two terms on implementation: feedback quality and identification (*B* = *0.06, SE* = *0.37, Exp(B)* = *1.06, p* = *0.87)*, feedback quality and explanation (*B* = *0.06, SE* = *0.35, Exp(B)* = *1.06, p* = *0.87)*.

Table 7
Two-level logistic regression analysis of the effects of quality and frequency on the relationship between features and implementation.

Predictors	Feedback features (Model 1b)		Feedback quality and features (Model 2b)		Feedback frequency, and features (Model 3b)		Feedback quality, frequency, and features (Model 4)	
	B (SE)	*Odds ratio*	*B (SE)*	*Odds ratio*	*B (SE)*	*Odds ratio*	*B (SE)*	*Odds ratio*
Mean outcome in logits (γ_{00})	−0.83 (0.64)	0.44	−1.17 (0.65)	0.31	−0.97 (0.66)	0.38	−1.29 (0.67)	0.28
Feedback quality	–	–	0.55 (0.15)	1.74***	–	–	0.52 (0.15)	1.68***
Feedback frequency	–	–	–	–	0.77 (0.09)	2.15***	0.76 (0.09)	2.14***
Quality*frequency	–	–	–	–	–	–	–	–
Feedback features								
Mitigating praise	0.08 (0.15)	1.09	0.10 (0.15)	1.10	0.13 (0.15)	1.14	0.15 (0.15)	1.16
Identification	0.30 (0.17)	1.35	0.35 (0.17)	1.42*	0.32 (0.18)	1.38	0.37 (0.18)	1.45*
Explanation	0.42 (0.16)	1.52**	0.39 (0.16)	1.47*	0.45 (0.17)	1.56**	0.41 (0.17)	1.51*
Solution	0.11 (0.20)	1.12	0.09 (0.20)	1.10	0.30 (0.21)	1.35	0.29 (0.21)	1.33
Suggestion	0.31 (0.17)	1.36	0.26 (0.17)	1.30	0.20 (0.18)	1.22	0.15 (0.18)	1.17
Control variables								
Total implementable comments	0.06 (0.04)	1.06	0.06 (0.04)	1.06	0.05 (0.04)	1.05	0.05 (0.04)	1.05
Number of spelling errors	−0.13 (0.06)	0.87*	−0.14 (0.06)	0.87*	−0.15 (0.06)	0.86*	−0.15 (0.07)	0.86*
Author's first draft quality	−0.28 (0.24)	0.76	−0.24 (0.24)	0.78	−0.26 (0.24)	0.77	−0.23 (0.24)	0.79
Gender (F = 1)	0.57 (0.41)	1.77	0.58 (0.41)	1.79	0.64 (0.42)	1.90	0.65 (0.42)	1.92
Age	0.02 (0.39)	1.02	0.01 (0.39)	1.01	−0.03 (0.39)	0.97	−0.04 (0.39)	0.96
Schools 2–6 (reference: School 1)	ns.		ns.		ns.		ns.	
Model fit statistics								
Deviance	−838.9		−832.1		−797.5		−791.8	
AIC	1711.8		1700.3		1631		1621.7	

Note. N = *1,582.* "-" means that the variables were not included in the model.
**p* < .05, ***p* < .01, ****p* < .001, *ns.*: not significant.

mediator (e.g., in the case of comment quality).

The most surprising result was the non-interaction between quality and quantity: students were also more likely to implement low-quality comments if repeated. But the follow-up analysis revealed an important countervailing force within multipeer review: students most often appeared to implement low-quality feedback when the comment overlapped with good quality feedback. In other words, students appeared to use good quality feedback to filter other feedback before adopting it. These findings present a generally positive picture of multipeer feedback: students discount low quality feedback and they appear to benefit from overlap in peer comments.

The common existence of mixed quality feedback across comments raises a new question regarding why students often focus on similar concerns but then give mixed quality advice. Further research should be conducted to examine the source of the mixed reviewing behaviors. For example, students might be more likely to respond to the received comments that share the same topics as the comments they provide as reviewers (Hovardas et al., 2014; Zhang et al., 2017). The feedback received and provided by students can be compared to examine the effects of the overlap between received and provided feedback on revisions and learning. In addition, whether students always provide comments on areas that they tend to have already mastered could be investigated in future research.

The negative correlation between identification and feedback quality (see Table 5) might help explain why students implemented low-quality peer feedback that was mentioned more than once. Students might have difficulty clearly identifying the most central, difficult problems. In addition, it is not yet clear whether students focused on the higher quality parts of mixed quality feedback. It is still possible that students might have implemented low-quality feedback mentioned more than once because of conformity and in-group influence (Stallen et al., 2013). In order to fit in with the group, students may conform to

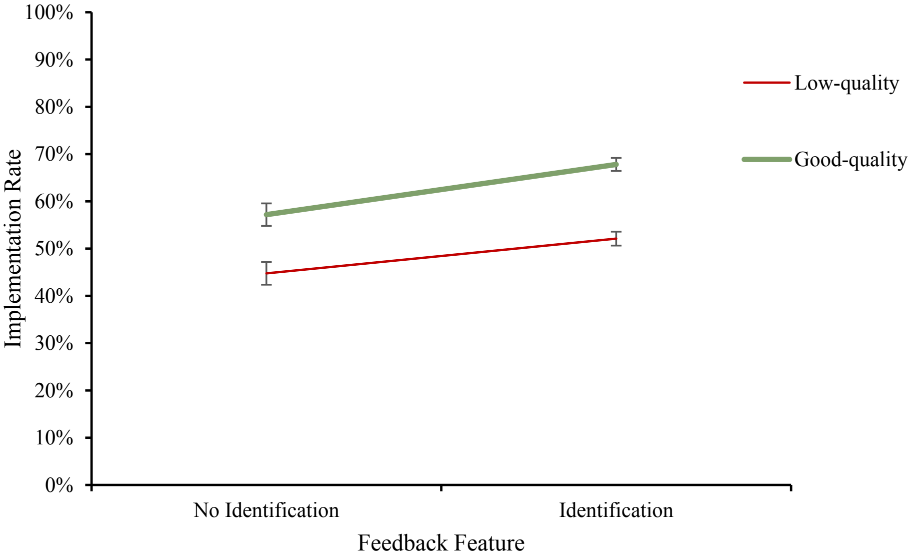

Fig. 3. Mean feedback implementation rates (with SE bars) as function of identification and feedback quality, for high-level comments.

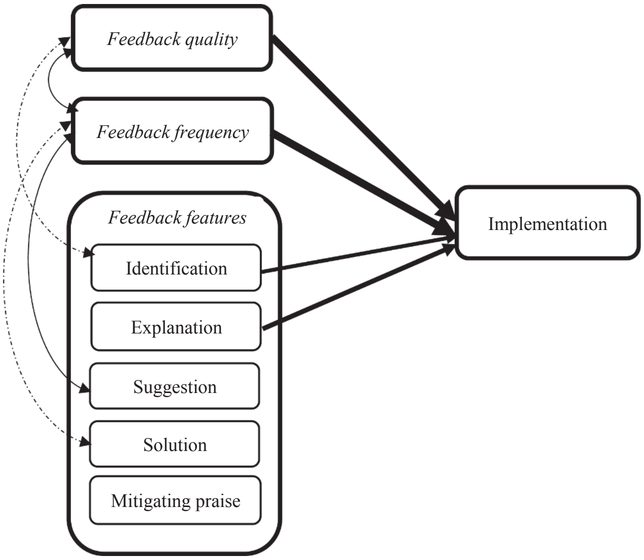

Fig. 4. Revised feedback-to-implementation model. Line thickness corresponds with statistical strength of relationship in the regression models. Curved lines with double-headed arrows reflect that the two variables are conceptualized as correlated rather than causal.

the behavior or the expectation of other group members, and follow the advice provided by their peers. Research done with larger samples could examine whether this pattern differs from good to poor writers, or across problems of different difficulties. It might be challenging for poor writers to distinguish accurate and inaccurate feedback on complicated high-level problems.

RQ 2: Does controlling for feedback quality and feedback frequency change the observed relationships between feedback features and feedback implementation?

When predicting feedback implementation, explanation was found to be a robust predictor, which is in line with previous research (Gielen et al., 2010; Wu & Schunn, 2020). The previously offered rationale for this effect is that explanatory feedback appears to help students understand received feedback and thus improves implementation rates. Explanations have also been found to be more important than peer feedback accuracy for improving students' writing performance (Gielen et al., 2010). A new rationale is supported in the current study: comments with explanations are also more likely to be implemented when feedback quality and frequency are held constant. Thus, including explanatory information in comments should be emphasized in peer review training.

When feedback quality was included, identification became a significant predictor of implementation. Prior studies also did not find the significant direct relationships of identification with implementation (e.g., Wu & Schunn, 2020) perhaps because identification alone does not help students understand problems, an important step towards implementation. The only prior work finding a positive effect of identification was at the more aggregate learner level and with draft quality rather than implementation (e.g., Lu & Law, 2012). In the present study, identification was correlated with implementation perhaps because students' attention could be drawn to the problems explicitly identified. Due to the confounding negative correlation between identification and quality (see Table 5), future research could further investigate the relationships between identification and quality.

Different from some prior research (e.g., Leijen, 2017; Nelson & Schunn, 2009), we did not find that suggestions and solutions predicted implementation on the comment level, although others have found no effect of suggestions on implementation (Patchan et al., 2016; Wu & Schunn, 2020). As students receive feedback from multiple reviewers on the same rubrics, it is likely that two comments focus on the same or similar aspects of the text. When students use good-quality feedback to filter low-quality feedback, or when students decide to use repeated feedback, suggestions and solutions appear to play less significant roles in revisions because the suggestions or solutions embedded in good comments could be used to solve the problems identified by low-quality comments. If this were true, identification and explanation would be more important than suggestions and solutions in that identification and explanation help students know what and where the problems are, and suggestions and solutions can be shared.

Additional predictors of implementation

Only one comment-level control variable significantly predicted implementation, i.e., spelling errors in the comments. The negative correlation between spelling errors and implementation may have occurred because spelling errors distracted receivers from reading so that they could not focus on the message of the feedback (Martin & Ranson, 1990). Since spelling errors are not necessarily damaging to the meaning-level, future research should examine this general effect in greater detail.

In addition, girls appeared to be more likely to act on received feedback than boys when feedback was mentioned more frequently. Girls and boys not only perform differently in writing ability (Reilly et al., 2019), but also appear to respond to peer feedback in different ways, especially when feedback is repeated. This difference in response to feedback may further widen performance differences by gender because revision work can improve students' writing ability (Patchan et al., 2016; Topping, 1998; Winstone et al., 2017). Thus, it may be that boys should get more support to become more active in peer feedback and revisions.

6. Implications for practice

In the 21st century, learning is becoming more of a participatory and collaborative activity (Barab et al., 2001; Hovardas et al., 2014). As part of that more general trend, multipeer feedback may become an indispensable feature of writing instruction (Schunn & Wu, 2019). To maximize the benefit of peer feedback on student writing performance, students should address received feedback in revisions (Winstone et al., 2017). Prior research on peer feedback has examined the influence of feedback features or general feedback frequency on revised draft quality (e.g., Gao et al., 2019; Leijen, 2017; Lu & Law, 2012; Tseng & Tsai, 2007), but the impact of feedback quality has been ignored. The current findings offer several implications for peer feedback practices within writing instruction.

First, based on the positive correlation between feedback frequency and implementation even for lower quality feedback, teachers could "assign students to give feedback on areas demonstrating students' strengths" (Chong, 2017: 22) so that it is more likely that students exclusively receive quality feedback. For example, students who are good at explaining strategies in their writing could be allocated to give feedback on this dimension to those who are poor at explaining strategies. Future research can compare the quality of feedback provided by matched peer review and non-matched peer review based upon this more topic-specific matching. Similarly, given the possible effect of ingroup influence and conformity, focusing reviewing on more narrow topics or arranging for more peer reviewers will increase the likelihood of receiving multiple comments on a key topic, which would then improve implementation rates. At the same time, instructors should take measures to mitigate conformity and encourage students to think critically. For example, example comments that are of low-quality but repeated can be presented to show that a group member may have a correct answer that the group does not have (Fender & Stickney, 2017). Receiving does not mean implementing, and declining wrong answers is also a valuable learning process.

Second, because feedback quality is also significant in predicting implementation, students should be invited to discuss feedback quality criteria so that they could have better understanding of what high quality feedback involves, which in turn helps them identify problems clearly, correctly, and consistently. In terms of the importance of feedback quality, some strategies can be adopted to improve feedback quality. For example, according to Gielen and De Wever (2015b), including guiding questions (e.g., what would you revise?) can lead to including more information in peer feedback.

Third, some techniques, such as Natural Language Processing and machine learning (Nguyen et al., 2016), could be used to detect feedback features (e.g., explanation) and comment spelling errors, and immediately remind students to add explanations or fix spelling errors during the peer review process.

Fourth, more support should be provided to boys when using peer feedback in the classes. Since differences between boys and girls are often found in their writing performance (Reilly et al., 2019), the differences observed here in rates of responding to feedback might be related to differences in their writing proficiency. However, they may also be related to prior experiences with peer feedback or motivational differences. Tentatively, teachers should consider giving more encouragement to address feedback or perhaps model strategies for responding to feedback.

7. Limitations and considerations for future research

There are a few limitations to the present study. First, the results are primarily quantitative and could be complemented with more in-depth qualitative data collection such as with interviews or reflective journals. Although post-hoc interpretations of why students implemented the repeated low-quality feedback were provided in the present study, qualitative data elicited from students would provide more in-depth information related to their revision behaviors.

Second, the factors influencing feedback implementation related to feedback on low-level aspects of writing (e.g., spelling and grammar) were not examined in the present study. Analyzing peer feedback on low-level aspects might surface different findings because of different cognitive loads involved (Liou & Peng, 2009) or different base-rates of comment features (Wu & Schunn, 2020). Research with more fine-grained distinctions of high-level problems (e.g., thesis, organization, explanation) might also provide more practical guidance to students and instructors.

Third, the limited sample size of the study prevented us from conducting mediation analysis of the effects of feedback features with feedback quality as a mediator. That is, it is possible that feedback features were associated with implementation because of differences in feedback quality. Future research could use larger sample sizes to test the relationship between feedback features, quality, and implementation, which would provide valuable information on how to improve feedback quality and support peer feedback.

Third, the correlational research design of the study prevents strong causal conclusions, although the main plausible confounds suggested by prior research were ruled out. In the current study, we have narrowed the focus of which factors are worth testing and controlling in experimental studies. Finally, other factors (e.g., conformity) that might explain feedback implementation, especially repeated low-quality feedback, can be included. For example, researchers (e.g., Bond & Smith, 1996; Fender & Stickney, 2017; Stallen et al., 2013) found that culture background, competence, and personality can contribute to conformity in group work. Including these learner factors as covariates may help identify factors underlying changes in implementation.

8. Conclusions

We examined the relationships of feedback quality and frequency to implementation in a secondary AP writing course in the study. We also investigated whether the inclusion of feedback quality and frequency changed the effects of feedback features on implementation. Both qualitative and quantitative methods were used to analyze the data and answer the two major research questions. Results indicated that feedback quality and frequency were two strong predictors with robust effects to predict implementation. Students implemented high-quality, medium-quality, and low-quality comments as the comments were repeated, but much higher implementation rates were associated with higher feedback quality. Further, students implemented the low-quality comments that were mentioned more than once primarily because the comments overlapped good comments. In terms of the relationship of feedback features to implementation, identification became significant when feedback quality was included. The effects of another feature, explanation, were robust.

The current study provided important new evidence to encourage the use of multipeer review: the relatively rare occurrence of low-quality feedback, the positive effects of feedback quality on implementation, and the large effects of feedback frequency on implementation. Teachers are therefore encouraged to arrange for multipeer feedback in writing instruction to help students revise. In conducting multipeer feedback, teachers could try matching students according to their strengths and weaknesses, asking students to include explicit identification and detailed explanation, and giving more encouragement to boys to make use of the feedback received.

Declaration of Competing Interest

The second author is a co-inventor of the peer review system used in the study.

Funding

The original data collection was funded by a grant from the US Institute of Education Sciences.

Appendix A. Peer review rubrics

Thesis Did the author include a clear, specific thesis in his or her introduction?

7 - The author's introduction includes a clear, specific thesis statement that connects Louv's rhetorical strategies with the argument he is making about the separation between people and nature.

6 - 6

5 - The author's introduction includes a thesis, but the thesis does not make a specific or clear connection between Louv's rhetorical strategies and his argument about the separation between people and nature.

4 - 4

3 - The author's introduction includes a thesis, but the thesis is overly general or simply a restatement of the essay prompt.

2 - 2

1 - The author did not include a thesis in his or her introduction.

Argument Did the author accurately describe Louv's argument about the separation between people and nature?

7 - The author accurately describes all of Louv's argument.

6 - 6

5 - The author accurately describes most of Louv's argument.

4 - 4

3 - In the majority of the essay, the author misunderstands Louv's argument.

2 - 2

1 - The author does not address Louv's argument and instead writes about his or her own argument about the separation between people and nature.

Rhetorical strategies What rhetorical strategies did the author analyze in his or her essay?

7 - The author analyses multiple, subtle rhetorical strategies that Louv uses accurately (such as appeal to a common cause, evoking nostalgia, or other sophisticated strategies).

6 - 6

5 - The author analyses three or more obvious rhetorical strategies that Louv uses (such as using rhetorical questions, anecdotes, or other obvious strategies).

4 - 4

3 - The author analyses only 1–2 obvious rhetorical strategies that Louv uses (such as rhetorical questions) or misunderstands Louv's strategies.

2 - 2

1 - The author didn't write about Louv's rhetorical strategies (instead discussed a different topic, connected to personal experience, or just summarized Louv's piece).

Evidence for claims How strong is the textual evidence for each claim about Louv's rhetorical strategies?

7 - Every claim has accurate evidence for all important aspects of the claim. Most evidence is conveyed through direct quotes.

6 - 6

5 - 5-Every claim has evidence, but some of the evidence is not accurate or not complete. Some evidence is conveyed through direct quotes.

4 - 4

3 - Several claims are missing evidence, or most of the evidence is not accurate. Little or no evidence is conveyed through direct quotes.

2 - 2

1 - No evidence is provided for any of the claims.

Explaining evidence Are the explanations of the textual evidence logical and thorough?

7 - Explanations of all the evidence provided are thorough, logical and connected to the essay's thesis.

6 - 6

5 - Explanations are sufficient, but not always thorough, logical, and clearly connected to the essay's thesis.

4 - 4

3 - Explanations are simplistic, sometimes absent, or not clearly connected to the essay's thesis.

2 - 2

1 - Explanations are missing or unrelated to the prompt (such as based in personal experience).

Organization Did the author organize his or her essay logically and clearly?

7 - The essay has a clear organization with a logical progression of ideas and body paragraphs that are each focused on a single argument that connects back to the thesis.

6 - 6

5 - The essay has a clear organization and progression of ideas, but the body paragraphs may sometimes be unfocused or not clearly connected to the thesis. The organization may be simplistic with formulaic transitions and a list-like progression of ideas.

4 - 4

3 - The organization of the essay is difficult to follow in many places due to jumps in logic, lack of transitions, repetition, and lack of focused body paragraphs that connect to the thesis.

2 - 2

1 - The essay is very disorganized with most ideas presented in random, repetitive, or illogical ways that make the author's argument and its connection to a thesis very difficult to understand.

Control of language How appropriate are the writing style and vocabulary for an academic essay?

7 - Mature, sophisticated prose style, using specific academic terminology (such as pathos and ethos) and control of language.

6 - 6

5 - Clear prose style with few lapses in academic word choice.

4 - 4

3 - The prose generally conveys the writer's ideas but is inconsistent in controlling the elements of effective writing, such as academic word choice.

2 - 2

1 - Simplistic style and vocabulary.

Conventions How well does the paper follow the conventions (grammar, punctuation, and spelling) of Standard Written English?

7 - The paper follows the conventions of Standard Written English very well with very few or no errors.

6 - 6

5 - The paper mostly follows the conventions of Standard Written English, but has about 1–2 error per paragraph. The errors don't interfere with your understanding the writer's ideas.

4 - 4

3 - The paper does not consistently follow the conventions of Standard Written English and may include up to 3–5 errors per paragraph. In places, the errors make it hard to understand the writer's ideas.

2 - 2

1 - In many sentences, the paper does not follow the conventions of Standard Written English. The errors make it very difficult to understand the write's ideas in many places.

Appendix B

Coding scheme with examples of high-level feedback topics, organized by writing dimension

Dimensions	Feedback topics	Feedback examples
Thesis	1. There is no thesis in the introduction.	*There was no thesis in the passage.*
	2. The thesis is overly general or simply a restatement of the essay prompt.	*While the essay was organized, from reading the thesis it is unclear where you may be going with the essay.*
	3. The thesis does not make a specific or clear connection between rhetorical strategies and the author's argument.	*I feel like this is a thesis, but you really didn't mention the problem that Louv is talking about, which is the separation between nature and technology.*
	4. There is no background information.	*The writer did not offer any background information on the topic in which he or she was writing about.*
	5. The thesis does not align with body paragraphs.	*Your thesis statement mentions four things: Anecdotes, imagery, ethos, and pathos. You made sure to touch on anecdotes and imagery in your body paragraphs, but barely talk about pathos and ethos.*
	6. Others (i.e., problems that are not covered by the above five categories)	*Try to avoid using quotes from the text in your thesis paragraph.*
Organization	1. Organization is not clear (e.g., ideas are presented in random)	*Your essay has no pattern; it was hard for me to read.*
	2. Body paragraphs are unfocused.	*For the metaphor and repetition paragraph, it was a bit all over the place. You started out analyzing her repetition and then you go on to talk about a metaphor; and in your last sentence, you went back to talking about repetition.*
	3. Body paragraphs are not clearly connected to the thesis.	*Your paper was very well organized in the AP format. Your conclusion cohesively wrapped everything together but just don't forget to do that in the middle paragraphs. Make sure to go back to the thesis at hand.*
	4. Formulaic or lack of transitions	*Try and use transitions at the beginning of your body paragraphs.*
	5. Others (i.e., problems that are not covered by the above four categories)	*I was a little confused by the last paragraph. I couldn't exactly tell if the last paragraph was a body paragraph or a conclusion.*
Argument	1. The argument is not addressed.	*Every rhetorical strategy fell back on either technology, or nature, which was really the reason for writing. But it seemed as if you were only stating rhetorical strategies and not supporting Louv's true argument.*
	2. The argument is misunderstood.	*The author says Louv believes that an overlap of nature and technology is unacceptable. Louv doesn't believe that it is unacceptable, he just believes that the world will change for the worse because of it.*
	3. The argument is vague/incomplete.	*I understood the point you were trying to make in analyzing your rhetorical strategies, but overall your argument seems weak and vague. Try to go more in depth with your thinking, and you will have a stronger and more persuasive essay.*
Strategies	1. There is no rhetorical strategy.	*You didn't have any rhetorical strategies, you just explained and summarized his argument.*
	2. Analyze only obvious rhetorical strategies.	*Your rhetorical strategies were too general. There were many to choose from in the essay. Try finding stronger strategies that you can support more thoroughly.*
	3. Rhetorical strategies are not enough.	*You could improve your essay by including a third body paragraph analyzing a technique that is very different to logos and pathos.*
	4. Rhetorical strategies are misunderstood.	*First body paragraph: "business roles" is not a rhetorical device.*
	5. Rhetorical strategies are vague.	*The author should mention what the rhetorical devices were next time in order to make it clearer for the reader.*
	6. Too many rhetorical strategies	*You listed multiple strategies and you only explained one or two of them.*
Evidence	1. Evidence is not accurate.	*You do an amazing job of explaining your real-world examples. In your third body paragraph, however, you add too much biased information for your point.*
	2. Evidence is not complete.	*The author did not provide readers with complete quotes and constantly use single worded quotes that became repetitive and unneeded.*
	3. Evidence is not conveyed through direct quotes.	*You need quotes to support your topic sentences. Quotes make your essay better.*
	4. Evidence is not enough.	*More evidence is needed to back up your claims. There were direct quotes but not enough to fully express your ideas and claims.*
	5. Too many pieces of evidence	*You have lots of evidence for your claims, in fact maybe too much. Most of your paragraphs are overpowered by quotes leaving no space for analysis.*
Explanation	1. Evidence/argument is not explained.	*In paragraph 2, you provide some of Louv's rhetorical questions. But you do not explain what they make the reader think about. You must describe how the questions play a role in the readers minds.*
	2. Explanations of evidence/argument are simplistic, or not clearly connected to the argument.	*Explanation is done, but could go deeper to properly analyze the document.*

Appendix C

Examples of qualitative analysis of low-quality comments that were repeated

Example A:

***(Original Para 2) The first strategy that Louv used that jumped out of the page was his use of sarcasm**. "Yes, we'll say, it's true. We actually looked out the car window." (61) Obviously when people are in the car, they'll occasionally look out the window, but by his sarcasm, it makes the reader feel that technology will become so prevalent in cars that kids of the future won't even glance out the window. This makes a very powerful point since the reader probably laughed about this at first, but then realized the deeper meaning by it.*

***(Revised Para 2) Louv uses sarcasm as a rhetorical device to get his argument across**. "Yes, we'll say, it's true. We actually looked out the car window." (61) Obviously when people are in the car, they'll occasionally look out the window, but by his sarcasm, it makes the reader feel that technology will become so prevalent in cars that kids of the future won't even glance out the window. This makes a very powerful point since the reader probably laughed about this at first, but then realized the deeper meaning by it.*

***Feedback 1**: The organization of the essay was good, but the introductions for each paragraph could be better. You either did not use one, or you used a very basic one. Use introduction words that fit perfectly with what that paragraph is about. For example, in paragraph two you could say "to excitingly start off…" since the paragraph is about sarcasm. (Formulaic or lack of transitions, Mixed, Implemented)*

***Feedback 2**: Don't use cookie-cutter transitions like "first", "next", and "also". Use full sentences to make transitions into paragraphs like "Louv uses sarcasm as a rhetorical device to get his argument across" or something like that. (Formulaic or lack of transitions, Medium-quality, Implemented)*

Example B:

(Original Thesis) *Louv uses personal experiences as examples to show how the importance of nature has changed throughout generations. Louv also develops his argument by talking about how culture has changed due to the growing ignorance of nature and brings about a feeling of nostalgia in the reader.*

(Revised Thesis) *As generations have passed, nature has become less and less important in culture and sometimes "isn't even worth looking at" because of this kind of misuse. Through personal experiences and diction that evokes emotional appeals to the reader, Louv fuels his argument.*

Feedback 1: *They didn't have much of a thesis. (No thesis, Incorrect, Implemented)*

Feedback 2: *Was not a thesis. (No thesis, Incorrect, Implemented)*

References

Barab, S. A., Hay, K. E., Barnett, M. G., & Squire, K. (2001). Constructing virtual worlds: Tracing the historical development of learner practices/understandings. *Cognition and Instruction, 19*(1), 47–94. https://doi.org/10.1207/S1532690XCI1901_2.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119*(1), 111–137. https://doi.org/10.1037/0033-2909.119.1.111.

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

Cho, K., Chung, T. R., King, W. R., & Schunn, C. (2008). Peer-based computer-supported knowledge refinement: An empirical investigation. *Communications of the ACM, 51*(3), 83–88. https://doi.org/10.1145/1325555.1325571.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*, 328–338. https://doi.org/10.1016/j.learninstruc.2009.08.006.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73–84. https://doi.org/10.1037/a0021950.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education, 48*(3), 409–426. https://doi.org/10.1016/j.compedu.2005.02.004.

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

Chong, I. (2017). How students' ability levels influence the relevance and accuracy of their feedback to peers: A case study. *Assessing Writing, 31*, 13–23. https://doi.org/10.1016/j.asw.2016.07.002.

College Board. (2018). Program summary report. Retrieved from https://securemedia.collegeboard.org/digitalServices/pdf/research/2018/Program-Summary-Report-2018.pdf.

Fender, C. M., & Stickney, L. T. (2017). When two heads aren't better than one: Conformity in a group activity. *Management Teaching Review, 2*(1), 35–46. https://doi.org/10.1177/2379298116676596.

Gao, Y., Schunn, C. D., & Yu, Q. (2019). The alignment of written peer feedback with draft problems and its impact on revision in peer assessment. *Assessment & Evaluation in Higher Education, 44*(2), 294–308. https://doi.org/10.1080/02602938.2018.1499075.

Ge, Z. (2019). Investigating the effect of real-time multi-peer feedback with the use of a web-based polling software on e-learners' learning performance. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2019.1643743.

Gielen, M., & De Wever, B. (2015a). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning, 31*, 435–449. https://doi.org/10.1111/jcal.12096.

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

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

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education, 40*(1), 151–164. https://doi.org/10.1080/02602938.2014.893418.

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education, 71*, 133–152. https://doi.org/10.1016/j.compedu.2013.09.019.

Hughes, G. D. (2012). Teacher retention: Teacher characteristics, school characteristics, organizational characteristics, and teacher efficacy. *The Journal of Educational Research, 105*(4), 245–255. https://doi.org/10.1080/00220671.2011.584922.

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods, 18*(1), 3–20. https://doi.org/10.1177/1525822X05282260.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6.

Leijen, D. A. J. (2017). A novel approach to examine the impact of web-based peer review on the revisions of L2 writers. *Computers and Composition, 43*, 35–54. https://doi.org/10.1016/j.compcom.2016.11.005.

Lin, G. Y. (2018). Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education, 116*, 81–92. https://doi.org/10.1016/j.compedu.2017.08.010.

Liou, H. C., & Peng, Z. Y. (2009). Training effects on computer-mediated peer review. *System, 37*(3), 514–525. https://doi.org/10.1016/j.system.2009.01.005.

Lu, J. Y., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science, 40*, 257–275. https://doi.org/10.1007/s11251-011-9177-2.

Martin, C. L., & Ranson, D. E. (1990). Spelling skills of business students: An empirical investigation. *The Journal of Business Communication, 27*, 377–400.

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., et al. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health, 60*, 290–297. https://doi.org/10.1136/jech.2004.029454.

Narciss, S. (2008). *Handbook of research on educational communications and technology* (pp. 125–143). ((3rd ed.)). Erlbaum.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401. https://doi.org/10.1007/s11251-008-9053-x.

Nguyen, H., Xiong, W., & Litman, D. (2016). Instant feedback for increasing the presence of solutions in peer reviews. Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT). San Diego, CA.

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*, 381–405. https://doi.org/10.1007/s11251-012-9236-3.

Patchan, M. M., Schunn, C. D., & Clark, R. (2017). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education. 1–16*. https://doi.org/10.1080/03075079.2017.1320374.

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*(8), 1098–1120. https://doi.org/10.1037/edu0000103.

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11*, 289–303. https://doi.org/10.1007/s10459-005-3250-z.

Prinsen, F. R., Volman, M. L. L., Terwel, J., & Van den Eeden, P. (2009). Effects on participation of an experimental CSCL-programme to support elaboration: Do all students benefit? *Computers & Education, 52*(1), 113–125. https://doi.org/10.1016/j.compedu.2008.07.001.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445–458. https://doi.org/10.1037/amp0000356.

Schunn, C. D., Godley, A. J., & DiMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy, 60*(1), 13–23. https://doi.org/10.1002/jaal.525.

Schunn, C. D., & Wu, Y. (2019). The learning science of multi-peer feedback for EFL students. *Technology Enhanced Foreign Language Education*, 13–21.

Schunn, C. D., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English school. *Journal of Adolescent & Adult Literacy, 60*(1), 13–23. https://doi.org/10.1002/jaal.525.

Sluijsmans, D. M. A., Brand-Gruwel, S., & Van Merriënboer, J. J. G. (2002). Peer Assessment Training in Teacher Education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454. https://doi.org/10.1080/0260293022000009311.

Stallen, M., Smidts, A., & Sanfey, A. G. (2013). Peer influence: Neural mechanisms underlying in-group conformity. *Front. Hum. Neurosci, 7*, 50. https://doi.org/10.3389/fnhum.2013.00050.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*, 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276. https://doi.org/10.3102/00346543068003249.

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161–1174. https://doi.org/10.1016/j.compedu.2006.01.007.

US Department of Education (DOE). (2018, October 24). Improving basic programs operated by local educational agencies (Title I, Part A). https://www2.ed.gov/

programs/titleiparta/index.html.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Berg (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009.

Wang, W. (2014). Students' perceptions of rubric-referenced peer feedback on EFL writing: A longitudinal inquiry. *Assessing Writing, 19*, 80–96. https://doi.org/10.1016/j.asw.2013.11.008.

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

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*. https://doi.org/10.1016/j.cedpsych.2019.101826.

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

Zhang, F., Schunn, C. D., & Baikadi, A. (2017). Charting the routes to revision: An interplay of writing goals, peer comments, and self-reflections from peer reviews. *Instructional Science, 45*, 679–707. https://doi.org/10.1007/s11251-017-9420-6.