Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback

Yong Wu[a*]

Christian D. Schunn[b]

[a]School of Humanities, Beijing University of Posts and Telecommunications

Beijing, 100876, China

[b]Learning Research and Development Center, University of Pittsburgh

3420 Forbes Ave

Pittsburgh, PA 15260, USA

[*]Corresponding author: Yong Wu (wuyong@bupt.edu.cn)

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

Author Note

The data and analysis code to this article can be found at https://osf.io/ta3cu/.

Citation: Wu, Y. & Schunn, C. D. (In press). Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *Journal of Educational Psychology*

Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback

Abstract:

Although peer review has been widely used for formative assessment in writing instruction, there remain concerns about whether assessors are at a sufficient writing performance level that would allow them to identify major problems in the reviewed work and provide helpful feedback to improve draft quality. Little empirical research has examined how assessor writing performance specifically influences problem identification accuracy and helpfulness of feedback, nor has it acknowledged different grain sizes of assessor performance. Assessor writing performance at different grain sizes (i.e., performance at the levels of genre, dimension of a genre, and specific problem topic) was assessed alongside problem identification accuracy and feedback helpfulness in 234 high school students who participated in an anonymous multi-peer review in a secondary writing course in the United States. A correlation analysis showed that assessor performance levels on specific problem topics were meaningfully separable, thereby allowing for consideration of the effects of assessor performance at genre, dimension, and topic levels. Multiple regression results indicated that assessor writing performance was unrelated to problem identification accuracy at any grain size. Therefore, scaffolds in the reviewing process appear sufficient to support problem identification accuracy. However, assessor writing performance, particularly on specific dimensions and specific topics, consistently predicted helpfulness of feedback, even though lower performing assessors rarely produce incorrect advice. Theoretical and practical implications of the findings are discussed.

Key words: Peer review, Assessor writing performance, Identification accuracy, Helpfulness of feedback, Feedback features

Educational Impact and Implications Statement

1. Assessor writing performance is conceptualized in terms of different grain sizes: overall writing performance, genre writing performance, writing performance on a cluster of dimensions, dimensional writing performance, and topic writing performance.
2. An assessor who is at a given writing performance level at a large grain size may often be at varying performance levels at more fine-grain sizes.
3. Assessors at both high and low writing performance levels, at any grain size, are equally likely to identify problems accurately.
4. Assessors with high performance on specific dimensions and topics are more likely to provide helpful feedback on those dimensions and topics.

1. Introduction

Peer review is defined as information provided by learners of similar status to peers to improve their work. It has been widely used in both K-12 and tertiary education (Li et al., 2020; Shute, 2008; Topping, 1998; Zheng at al., 2020), and across many different disciplines, including writing instruction, science, engineering, physical education, and music. Peer review is commonly used as an alternative or supplement to teacher feedback for purposes of formative assessment or summative assessment because of its immediacy, frequency, amount, individualization, and learning benefits (DiPardo & Freedman, 1988; Gielen et al., 2010; Hovardas et al., 2014; Topping, 1998; Zheng et al., 2020). It can include verbal or written comments that provide diagnostic information and rubric-based scores that indicate students' current performance levels. Peer feedback specifically refers to the comments provided by peer assessors, and these peer comments are considered particularly helpful for promoting learning when they include a detailed diagnosis and suggestions for how to improve (Huisman et al., 2017; Wooley et al., 2008; Zong et al., 2021).

Conceptually, peer feedback is a reciprocal process in which students not only receive and act on received feedback as assessees, but also practice their evaluative judgement as assessors (Carless & Boud, 2018; Wu & Schunn, 2020a). Students, as assessor or assessee, benefit from peer feedback (Double et al., 2020; Sadler & Good, 2006; Topping, 1998), especially when peer review is conducted using online peer assessment systems that include a number of supporting functions (e.g., anonymous peer feedback, back-evaluations, multi-peer assessors, well-structured and reader-friendly rubrics) for enhancing peer feedback quality (Double et al., 2020; Hovardas et al., 2014; Panadero et al., 2013; Strijbos et al., 2010; Wu & Schunn, 2020b). Receiving feedback enables assessees to better understand the problems in their submissions and think critically, whereas providing feedback helps assessors develop audience awareness, apply criteria when reviewing, appreciate advantages of others' work, practice revising skills by identifying problems and suggesting revisions, and reflecting on their own work.

However, despite the extensive research supporting peer feedback, instructors and students often have doubts about whether peer assessors have sufficient expertise to provide useful feedback (Wu, 2019; Wu & Schunn, 2020b; Yu & Lee, 2016). In particular, there are concerns about whether peer assessors, especially those with low performance in the domain, can identify the main problems in assessees' work or provide useful feedback to help assessees revise. Some research suggests that assessors with higher task performance are more likely to provide useful feedback (Huisman et al., 2017; Patchan et al., 2013), but other research finds limited effects of assessor abilities on review quality (e.g., Patchan & Schunn, 2016). Many studies find that peer feedback is generally helpful for student learning (Graham & Perin, 2007; Huisman et al., 2018; Liu & Carless, 2006; Shute, 2008; Zheng at al., 2020), but other research suggests feedback quality depends upon the difficulty of the problems to be identified, with some problems being well beyond what any of the peers can address (Gao et al., 2019). Concerns about the capabilities of some or all peer assessors can influence both instructors and students' willingness to make use of peer feedback (Kaufman & Schunn, 2011). Therefore, it is important for research to provide insights into when such concerns are well founded and when they are not.

In addition, the effects of assessor writing performance at different grain sizes on feedback provision has received very little attention. Conceptually, grain sizes refer to the nested aggregation levels (from macro to micro) for analyzing a student's writing performance level. The existing peer feedback literature has generally focused on assessor overall writing performance (e.g., Wu, 2019; Yu & Lee, 2016; Zhao, 2011). However, writing draws upon multiple levels of performance, such as general linguistic performance, writing performance within a particular genre (e.g., short stories vs. persuasive essays), and performance with particular writing moves (e.g., providing supporting evidence or transitions between paragraphs). Further, research on assessor writing performance at different grain sizes (e.g., specific problems in writing) remains scarce.

According to common writing process models (e.g., Flower et al., 1986), authors need to be able to detect problems and then repair problems in their own writing. Similarly, within the process of peer feedback, assessors need to identify problems and then provide information to help assessees solve the problems. Thus, identifying problems and providing helpful feedback are two different subprocesses and might require different knowledge and skills, although they are often considered together as feedback quality (e.g., Hovardas et al., 2014). Here we examine whether writing performance at fine-grained sizes are especially central to the accuracy of problem identification and helpfulness of constructive comments within peer feedback. We suggest that attention to grain-size of the assessor's writing performance can help to clarify when and in what ways peer feedback is likely to be accurate and useful.

2. Theoretical Background

2.1 Assessor Writing Performance at Different Grain Sizes

Assessment of performance in writing can occur at multiple grain sizes depending upon the purpose: overall writing performance across multiple genres (e.g., expository writing and argumentative writing) for university placement exams; a specific genre (e.g., argumentative writing vs. expository writing vs. short stories) for summative evaluation in a class; writing performance on a cluster of dimensions within a writing assignment (e.g., higher-level writing issues); or specific dimensions within a genre (e.g., thesis, evidence, explanation, organization, language) and specific kinds of errors/micro-competencies within a dimension (e.g., clarity of thesis statement, quantity of evidence, quality of explanation, transitions, writing style) for formative feedback (see Figure 1). Within the research question of whether students are sufficiently competent to diagnose issues and provide constructive comments on their peer's work, we suggest it is very important to consider the grain size of the assessor's performance, especially the finer grain-sizes that are most relevant to formative feedback. The left side of Figure 1 shows the full continuum of grain-sizes of assessor writing performance. Their potential connection to the assessor's feedback quality is shown on the right.

In the peer feedback literature, assessors' writing performance has typically been operationalized as language proficiency or overall writing performance. Overall writing performance is usually measured using a language proficiency test for studies conducted with foreign language or second language learners (e.g., Allen & Mills, 2014; Wu, 2019; Yu & Lee, 2016) or overall writing performance without any specific genre focus for studies with native language speakers (e.g., Patchan & Schunn, 2015). However, even low English proficiency assessors have been found to provide feedback that is generally implemented by assessees in revisions (Yu & Lee, 2016). Similarly, assessors' English proficiency appears not to influence feedback quality when defined as to what degree peer feedback improved the original text

(Wu, 2019). But on some aspects of peer feedback quality, research studies focusing on assessors' general language proficiency have yielded mixed results. For example, two studies found that assessors' overall writing/English proficiency appeared not to influence feedback amount (Patchan & Schunn, 2016; Wu, 2019), but one study using a general English aptitude assessment observed more feedback given by higher English proficiency assessors than by lower English proficiency assessors (Allen & Mills, 2014).

Although not described by the study authors in this way, some peer feedback studies actually focused on the effects of assessor writing performance within a specific genre because they only measured writing performance within a particular genre (e.g., Chong, 2017; Huisman et al., 2017). For example, Huisman et al. (2017) observed that assessors with higher writing performance in the genre being peer reviewed provided more content-related feedback. This focus on genre performance could be taken as a wise focus because students' writing performance often varies according to task types (Keller et al., 2020). As noted by Tai et al. (2018: 472) "Evaluative judgement is domain specific: one develops expertise pertaining to a specific subject or disciplinary area, from which decisions about quality of work are made." Zhao (2011) found that deficits in genre knowledge often limited assessors' competence to provide feedback to English poems. As an important factor that shapes feedback process, genre knowledge helps students provide quality feedback. Further, if students are familiar with genre features, the cognitive demand in reviewing genre texts will likely be reduced (Wu, 2019).

Assessor writing performance can also be conceptualized in terms of a cluster of dimensions within a writing task (e.g., high-level vs. low-level writing issues). High-level and low-level aspects of writing are frequently analyzed separately because different patterns of results in peer feedback studies have been observed between the two levels (Liou & Peng, 2009). Previous research has found positive correlations between assessors writing performance at the grainsize of a cluster of dimensions and the quality of the feedback they provide on those dimensions. For example, Wu and Schunn (2020b) observed a positive relationship between assessors' writing performance on high-level/low-level aspects and the number of high-level/low-level comments. Chong (2017) found that students with higher levels of writing ability in terms of content development were able to provide more high quality feedback on content-related problems, and students who possessed high levels of language ability were able to identify grammatical errors more accurately.

Writing performance of peer assessors could also be operationalized as specific dimensional performance within a particular genre because different knowledge and skills are involved in different dimensions within a genre (e.g., thesis quality vs. evidence quality vs. explanation quality in argumentative writing). Although genre-based pedagogical approaches have been practiced in writing instruction for many years, genre-focused peer feedback has seldom been examined in writing instruction from the perspective of providing feedback (Yu & Lee, 2016). Yu (2021) investigated EFL learners' practice in providing genre-based peer feedback on theses/ dissertations, but he did not directly measure peer assessors' writing performance on different dimensions within thesis/dissertation writing.

No prior research has focused on assessor writing performance of feedback providers at very fine-grained levels (i.e., specific problems with a dimension of a particular genre), although a few studies have investigated peer feedback at the fine-grained level from the receiving side (e.g., Gao et al., 2019; Wu & Schunn, 2020b). It seems likely that poor performance on specific problems would prevent peers from providing helpful feedback on those specific problems (Zhu & Carless, 2018), but this assumption has not yet been investigated empirically.

For feedback to initiate revisions, it should contain accurate diagnosis information. In other words, assessors are expected to identify problems correctly. Further, feedback should include helpful information (e.g., explanations and suggestions) to persuade assessees to translate feedback into revisions (Patchan et al., 2016; Wu & Schunn, 2020a). Problem identification accuracy and helpfulness of feedback have been considered as major parts of feedback quality in prior research (e.g., Hovardas et al., 2014). We now turn to examining the potential role of assessor writing performance within identification accuracy and comment helpfulness.

2.2 Assessor Performance and Problem Identification Accuracy

Focusing first on accurate diagnosis, providing peer feedback to others obligates assessors to read essays for comprehension, identify problems, and diagnose them (Villamil & De Guerrero, 1996). These complicated processes pose a cognitive burden on assessors. Therefore, students with different writing performance are likely to execute the processes with different degrees of efficacy. When performing a cognitively complex task such as reviewing others' work, assessors with better writing performance have more cognitive resources to notice the problems in writing because they have stronger reading skills, better understand writing topics, and have more reader awareness (Van Steendam et al., 2010). Because peer feedback is generally conducted with a particular writing task, students with better genre writing performance are expected to be better at identifying problems in the writing task than those with poor performance, and thus may have higher accuracy rates in identifying problems during reviewing (Dijks et al., 2018; Huisman et al., 2017; Patchan & Schunn, 2015). However, people can notice problems in their own or others' work, suggesting that own writing performance is not strictly necessary for problem detection. Further, the rubrics provided to reviewers may provide sufficient scaffolding support to allow reviewers with weaker writing performance to nonetheless accurately detect problems.

Few studies have examined the relationships of assessors' dimensional writing performance and problem identification accuracy. Yu (2021) investigated postgraduate students' practices in providing genre-based peer feedback on theses/dissertations, e.g., "feedback on the purpose or functions of different parts of the thesis, feedback supporting students in meeting institutional requirements of what constitutes a thesis" (p. 43). He found that students could identify genre-related problems in spite of weaknesses in their own performance and knowledge of the genre, but sometimes provided incorrect and inappropriate genre-based feedback. The students compensated for their weaknesses in genre performance by using reference books, communicating with more capable peer students, and observing how their teachers provided feedback (Yu, 2021). However, assessors' genre writing performance was evaluated based on the interviews instead of via a performance measure.

At the most fine-grained size of specific-problems, Gao et al. (2019) investigated the relationship of assessor performance in terms of specific problems with the likelihood of providing feedback on those topics. They found that most of the peer feedback focused on less challenging problems; the specific problems that most authors had were rarely mentioned in the peer feedback. However, this research focused on patterns by particular problems across all students and did not directly examine the relationship of individual assessor performance on a problem and their feedback behaviors. Using a much larger dataset (previously examined to address other research questions in Wu & Schunn, 2020a), the current study coded assessor writing performance at different grain sizes, quality of reviewed documents, peer feedback identification accuracy and feedback helpfulness. Statistical analyses test the links between assessor writing performance with problem identification accuracy and feedback helpfulness.

2.3 Assessor Performance and Helpfulness of Feedback

Identifying problems is different from providing helpful feedback because the former only requires detection skills, but the latter also requires revision skills. Little attention has been given to the relationship between assessor genre writing performance and helpfulness of feedback. One indirect approach to addressing feedback helpfulness is to examine perceived helpfulness of feedback by the feedback recipient. One study found that assessors with higher writing performance were perceived as providing more helpful feedback (Dijks et al., 2018). However, perceived helpfulness is not necessarily the same as actual helpfulness of feedback. For example, assesses might be fooled by confident sounding comments, and assessors with higher writing performance might produce comments showing greater confidence. Another indirect approach is to focus on the comment features that persuade assessees to make revisions, such as suggestions and explanations (Gielen & De Wever, 2015a; Prins et al., 2006; Wu & Schunn, 2020a, c). A number of studies have found that assessors' genre writing performance (measured by the overall writing score of the preceding or current essay) was positively correlated with including suggestions (Chong, 2017; Huisman et al., 2017; Patchan & Schunn, 2015) and with including explanations (Chong, 2017; Huisman et al., 2017). At the level of dimensional aspects of writing, Chong (2017) found that assessors who were strong in dimensional aspects of writing (i.e., content-related and linguistic performance) were able to provide content-related feedback including more explanatory information and more accurate grammar feedback. However, little else is known about the relationships of dimensional or more fine-grained measures of assessor writing performance and presence of useful comment features such as suggestions and explanations.

Importantly, it is difficult to determine helpfulness of feedback by only considering the feedback features (e.g., inclusion of explanations or suggestions). First, quality of documents reviewed varies and assessors can focus on different aspects of different essays, including more and less problematic aspects of an essay (Hovardas et al., 2014). Second, assessors need to provide *meaningful* explanations and *useful* suggestions. Accordingly, some research has directly coded the helpfulness of feedback using experts (e.g., Wu & Schunn, 2020c), and that is the approach taken in the current study. Further, such research has found that explanations in peer feedback were not always helpful, possibly because the explanations were often inaccurate (Nelson & Schunn, 2009; Tseng & Tsai, 2007). Similarly, suggestions in peer feedback can sometimes fail to solve the problem or introduce new problems (Wu & Schunn, 2020c). However, it is possible that the peers tend to play to their own strengths and tend to provide comments when they feel confident, which is consistent with the observed low rates of incorrect feedback (Wu & Schunn, 2020c). But nonetheless, this would suggest a relationship between assessor performance, at least the micro-level, and the helpfulness of the suggestions they make. Overall, little research has been conducted on this topic, and it is very much an open empirical research question whether assessor writing performance (at any level) is related to helpfulness of peer feedback.

2.4 The Present Study and Research Questions

In sum, despite strong concerns about the ability of weaker writers to provide useful peer feedback, there is a lack of research that empirically tests the associations of assessor writing performance with problem identification accuracy and helpfulness of peer feedback. This research is critically needed to persuade instructors and students to use peer feedback, a pedagogical approach with strong empirical research. In addition, prior research has not directly considered the relative impact of assessor writing performance at different grain sizes. The purpose of this study was to investigate whether assessors' writing performance at different grain sizes predicted problem identification accuracy and helpfulness of peer feedback. The findings of the current study can therefore help identify the situations in which peer feedback can be fully trusted and the situations in which additional supports must be provided.

The proposed conceptual model of the hypothesized associations of assessor writing performance at different grain sizes, problem identification accuracy, and helpfulness of peer feedback is presented in Figure 1. Figure 2 presents the more specific model that is tested in the study. It assumes that assessor writing performance varies in separable ways at three grain sizes, e.g., genre performance (i.e., overall quality of the assessor's document in a particular assignment), dimensional performance (i.e., the quality of the assessor's document on the given dimension), topic performance (i.e., whether the assessor had a problem with the given topic within a dimension). Assessor performance at these three gain sizes are the key predictors. In terms of outcomes, problem identification accuracy refers to whether assessors correctly identified problems that actually exist. By contrast, helpfulness of feedback is defined in a way that speaks directly to the concerns of students and instructors: feedback is helpful if following the feedback will lead to improved quality in the revised draft.

Insert Figures 1 and 2

A number of control variables were included in the statistical models; the particular control variables varied by outcome variable (as shown by the arrows in Figure 2). Feedback dimensions, school type, document overall and dimensional quality were included as control variables when the outcome was problem identification accuracy. Document overall quality (i.e., overall quality of the assessed document) and document dimensional quality (i.e., the quality of the assessed document on the given dimension) were used to differentiate quality of the assessed document. Assessed document quality was included as control variables because previous research has found that the document quality influenced assessor' peer feedback performance (e.g., Lu & Law, 2012;

Wu & Schunn, 2020b). Feedback features, feedback topics, document overall and dimensional quality, and school type were included as control variables when the outcome was feedback helpfulness. Feedback features are treated as control variables because comments with more features might have different opportunities to be accurate or helpful. Document topic quality was not included in either model because an assessor only needed to detect a problem or provide helpful comments when the assessed document had a problem on that topic. There is also an assumed logical correlation between identification accuracy and helpfulness of feedback: If an assessor misses a problem, they will not provide any feedback. The definitions of the variables are presented in Table 1.

Insert Table 1

The following specific research questions are examined within this model. The first question is foundational to the second and third research questions (i.e., must be established first to make the other two questions meaningful).

1) Are different grain sizes of assessor writing performance empirically separable (i.e., at low enough correlations to allow analysis of their relative effects on identification and helpfulness)?
2) What are the relationships of assessor writing performance at different grain sizes with successful problem identification? We predicted stronger relationships at the micro level (e.g., assessor topic performance) and weaker relationships at the macro level (e.g., assessor genre performance).
3) What are the relationships of assessor writing performance at different grain sizes with usefulness of peer feedback? We predicted the same ordering, with stronger relationships at the micro level and weaker relationships at the macro level.

The study was conducted in the context of reciprocal online peer feedback in an advanced secondary writing course, but with a purposely broad sampling of school performance levels to address performance effects across an appropriately broad range of performance levels.

3. Method

3.1 Participants

Two hundred and thirty-four (234) secondary school students participated in the study. Participants were predominately female (59%; 2% did not report gender) and Caucasian (53%), followed by Asian (19%), African American (9%), and Hispanic/Latino (3%; 16% percent did not report race/ethnicity). The mean participant age was 17.2 years (*SD* = 0.5). Participants came from three secondary schools in the US. To broaden the applicability of the findings, two of the recruited schools (representing 43% of all the participants) were Title I schools, which receive additional financial support from the federal government to help students from low-income families (US DOE, 2018).

All participants were enrolled in Advanced Placement (AP) Language and Composition, a high school AP course aiming to be equivalent to a first-year writing course required by many universities in the US. Although there are clear systemic patterns in who is able to participate in AP courses, this course is one of the most widely taken AP courses in the US, with over 500,000 students taking the end-of-year exam in recent years (College Board, 2020). However, at the national level, the mean overall score of this course on the one to five scale has been consistently below a 3, the minimal level used to exempt students from taking the corresponding university course. The large number of students enrolled in the course as well as the low mean scores calls for more studies on how to help students improve their college-level writing ability.

Instructors were recruited for participation through email. Participating instructors were required to: have at least two years of experience teaching the AP writing course; currently teach at least two sections of that course; give a similar writing and peer assessment assignment at a common time during the school year; and use a common online system and common rubrics for the peer assessment aspect. All students in the course sections taught by the instructor participated in the study.

3.2 Materials

Peer assessment tool. Students used a shared online peer review system previously called SWoRD (Cho & Schunn, 2007; Schunn, 2016). Such tools can make large-scale peer review in classrooms easier and more efficient. The particular system, now called Peerceptiv, is currently used in a wide variety of courses by high schools and universities all over the world, and it has similar features to other commonly used systems like EduFlow, Kritik, PeerScholar, ELI Review, and MobiusSLIP. Instructors create writing assignments, specify rubrics, and set deadlines and other peer reviewing parameters in the system. Students submit their first drafts to the system before the deadline set by their instructor. Once submitted, the system allocates documents to multiple peers randomly and anonymously, and then it requires assessors to evaluate the drafts according to the rubrics provided by instructors. Students are generally given one week to provide feedback.

Writing and reviewing task. The writing task given in the peer review assignment involved an evidence-based rhetorical analysis that required students to read a one-page source passage and then write an essay discussing rhetorical strategies used by the author of the source passage. The rubrics of the reviewing task and accompanying commenting prompts (see Appendix A) included both high-level writing dimensions (i.e., thesis, argument, rhetorical strategies, evidence for claims, explaining evidence, organization) and low-level writing dimensions (i.e., use of language and writing conventions). The focus was predominantly on high-level dimensions, matching the advanced nature of the course. The genre was taken from the course curriculum, and the specific writing task was a released exam task from a prior year. The rubrics were based upon the expert scoring guide for the end-of-course exam, but adapted with student and teacher input to be student friendly (Schunn et al., 2016). Each student was required to review four texts within their teacher's classes in a double-blind fashion based upon the given rubrics that were shared across all schools.

3.3 Measures

Feedback topic of a comment, helpfulness of a feedback comment, feedback features included in a comment, student performance with respect to those specific feedback topics, and more general document quality representing student dimensional and genre performance were rated by trained coders (see Table 1 for the variables examined in the present study). In particular, feedback segmentation, feedback level, and feedback implementability were exhaustively coded by one writing instructor and two undergraduate research assistants. Feedback topics and feedback features were exhaustively coded by six undergraduate research assistants and one writing instructor. The helpfulness of feedback and overall document quality were exhaustively rated by four writing instructors. Student performance with respect to those specific feedback topics was coded by two writing instructors and two undergraduate research assistants. All the writing instructors had years of experience teaching English writing. All the undergraduate research assistants were English native speakers who were studying at a selective research university in the US. To ensure adequate interrater reliability (as assessed by having a Cohen's *Kappa* > .7), coders were iteratively trained to acceptable reliability thresholds, and then all data were exhaustively coded by at least two coders. All disagreements between the coders were flagged and resolved through discussion with a third coder present. Given the need to drop data at multiple steps in the process, a summary of each step in the feedback coding process along with the number of comments retained at each step is presented in Figure 3.

Insert Figure 3

3.3.1 Feedback Analysis

Feedback segmentation. The first step of feedback coding was to segment peer feedback into individual comments such that each comment focused on a single issue. Although students were given a reviewing interface that provided separate textboxes for each comment and were requested to focus each feedback comment on one problem, comments occasionally included more than one problem. For example, the single comment "*The thesis contains three main points and is fairly understandable but isn't written particularly well. It doesn't follow the preferred structure of a rhetorical analysis essay, and the several misspelled words are distracting.*" was segmented into two separate comments, one focused on the thesis (a high-level problem) and the other on spelling (a low-level problem). Feedback segmentation was conducted prior to feedback coding. However, during the coding process, a comment was sometimes further resegmented if it was found to contain more than one problem by coders. After segmentation (*Kappa* = 0.93), there were 7,964 comments in the dataset.

Feedback level. Each feedback comment was coded as being about a high-level or low-level writing issue (*Kappa* = 0.92, see examples in Appendix C). The goal was to separately analyze comments with a focus on high-level vs. low-level aspects of writing since different patterns of results have been observed by that distinction in previous studies on peer feedback (e.g., Liou & Peng, 2009; Wu & Schunn, 2020a, b). If a comment identified an issue related to the thesis, argument, rhetorical strategies, evidence for claims, explaining evidence, or organization, it was coded as a high-level comment. By contrast, low-level comments focused on the use of language (e.g., appropriate register or appropriate word choice) or writing conventions. A total of 5,994 comments were coded as high-level. The rubrics provided to assessors focus on six high-level dimensions and two low-level dimensions (see Appendix A). The assessors were required to provide at least one comment to each dimension. Therefore, high-level dimensions received far more comments than did low-level dimensions.

Low-level comments were excluded from further data analysis because they constituted the minority of the comments and because no specific low-level error occurred with high enough frequency to support analysis of performance on specific issues. For example, when looking at specific topics of low-level problems (e.g., wrong word choice, vague pronoun reference, run-on sentence, unnecessary or missing capitalization or comma), there were too few of each to conduct a statistical analysis. The relative amount of excluded low-level comments did not vary significantly across students.

Feedback implementability. Each high-level writing comment was then coded for whether it was implementable or not implementable (*Kappa* = 0.95). Following Patchan et al. (2016), if a comment could trigger a revision, it was considered implementable (e.g., *The thesis statement has clear points but is not constructed well.*). By contrast, if a comment only involved praise or only a summary statement, it was considered not implementable (e.g., *The author analyzed the rhetorical strategies very well.*). Approximately 2/3 (*N* = 4,022) of the high-level comments were coded as implementable. Only these revision-oriented comments were coded further given that helpfulness could not be coded within praise and summary comments. In addition, prior research suggests that revision-oriented comments are most important for revision and learning (Leijen, 2017) and that there were not enough summary or praise comments to examine accuracy within specific summary or praise topics.

Feedback topic. To strategically support a number of analytic functions, each implementable segmented comment was then coded for its topic (i.e., the specific problem identified in a comment). Because feedback topics sit at the intersection of the details of the specific assignment and the developmental trajectory of this set of students in terms of being able to notice particular problems, the coding scheme was iteratively developed from the dataset to identify common feedback topics. In the first round of feedback topic coding, 200 comments were selected randomly and inductively categorized by coders for underlying topics, i.e., what was the problem identified in a particular comment (e.g., vague thesis, lack of explanation). The preliminary coding scheme of common comment topics was then used to code another randomly-selected 300 comments. Some closely related topics were combined (e.g., general thesis & vague thesis) because of their low individual frequency or because of overlapping meaning that prevented reliable coding. The coding scheme was iteratively refined until all common topics were included and could be reliably distinguished. Not surprisingly, the final 27 comment topics each fell within one of the six reviewing prompt dimensions (e.g., not naming a rhetorical strategy, analyzing only obvious rhetorical strategies, and not analyzing enough rhetorical strategies were three specific comment topics that all fell under the strategies reviewing prompt). See Appendix B for the full set of 27 comment topics that were coded.

Then, topic frequency was calculated. The seven most common topics (found in 2,153 comments that were produced by 234 unique assessors) were: Quality of explanation (808 comments), Quantity of explanation (473 comments), Quantity of evidence (316 comments), Thesis clarity (168 comments), Body connected to thesis (154 comments), Connection in thesis

(130 comments), Transitions (104 comments) (see definitions and examples in Appendix B). These topics provided a sufficient number of comments per topic to form the basis of the topic-specific document quality coding (i.e., to reliably determine whether the assessor complained about a problem that in fact occurred in the reviewed document) and statistical analysis strategies that could control for topic effects (e.g., whether some topics were both rarely mentioned by lower performing assessors and happened to be often missed or incorrectly diagnosed). Thus, only these topics were used for analysis and further coding. The general discussion takes up the potential impact of focusing on this subset of the dataset since this was the largest strategic reduction in the overall size of the dataset.

Helpfulness of feedback. Comments within the seven most-common topics were coded for helpfulness of feedback (*Kappa* = 0.70). Helpfulness of feedback was defined to as the degree to which the comment could help the author improve quality of their document (Wu & Schunn, 2020c). Helpfulness of feedback was initially coded in terms of five categories (see Appendix C for definitions and examples): Feedback with large positive effect, Feedback with small positive effect, No effect feedback, Incorrect feedback (negative effect), and Mixed effect feedback (positive and non-positive effects when multiple recommendations are given). Since this coding only made sense when the coding topic was at least somewhat problematic in the document, it was only done for the 1,921 comments that were made on documents that had a problem with the given topic named in the comment (see Section 3.3.3 below).

Because several of the specific levels of feedback helpfulness were individually rare, the five quality categories were grouped for analysis into two levels: Helpful feedback (including Feedback with large positive effect and Feedback with small positive effect) and Unhelpful feedback (including No effect feedback, Incorrect feedback, and Mixed effect feedback). Helpful feedback could be used to improve quality of revised drafts, and Unhelpful feedback was unlikely to improve draft quality. However, early analyses using the more specific five levels of feedback helpfulness found similar results.

Feedback features. Prior research on peer feedback has found that several feedback features are associated with feedback quality (e.g., Wu & Schunn, 2020c) and revisions (e.g., Gao et al., 2019; Lu & Law, 2012; Nelson & Schunn, 2009; Schunn et al., 2016). To tease apart the effects of assessor performance on content accuracy of comments from confounding differences in feedback feature presence (e.g., were some students less likely to detect problems because of their writing performance level or the extent to which they produced substantive comments), five feedback features were coded (using the same N = 1,921 comments that were coded for helpfulness): Mitigating praise (*Kappa* = 0.87), Identification (*Kappa* = 0.78), Explanation (*Kappa* = 0.76), Suggestion (*Kappa* = 0.72), Solution (*Kappa* = 0.74) (see Appendix C for definitions and examples).

3.3.2 Student Writing Performance on Specific Feedback Topics

Students' first drafts were coded to assess performance on the seven most-frequent comment topics, which was used directly (i.e., to determine whether the document had an issue to identify) and indirectly (i.e., to determine whether the assessor showed high or low performance on the topic to understand whether that influenced the quality of the feedback they gave to others). Student performance on each specific topic was coded in terms of three categories: Competent, Ok, Poor. For example, performance on the topic "Quality of explanation" was determined to be Competent when they explained their opinions in a clear and logical way, Ok when they needed to improve on this topic to a certain degree, and Poor when their explanation was confusing.

The four coders first coded 20 randomly-selected drafts and discussed their coding to develop a stronger and similar understanding of topic performance coding. After that, they coded another 30 and then 50 randomly-selected drafts and discussed disagreements until the interrater reliability was acceptable: Quantity of explanation (*Kappa* = 0.70), Quality of explanation (*Kappa* = 0.69), Quantity of evidence (*Kappa* = 0.72), Body connected to thesis (*Kappa* = 0.70), Transitions (*Kappa* = 0.74), Thesis clarity (*Kappa* = 0.72), Connection in thesis (*Kappa* = 0.78). To ensure strong reliability within the final dataset given the modest reliability within some of the topics, all four coders coded all the remaining drafts. Ok and Poor were eventually combined for analysis to improve statistical power, although pilot analyses showed similar patterns of results when analyzed as a three-level variable.

3.3.3 Student Dimensional and Genre Writing Performance

In addition to coding for the seven high-frequency topic issues, more general document quality was also coded to measure student dimensional and genre writing performance. In particular, students' first drafts of the writing task were scored analytically by four trained writing instructors using the peer review rubrics (see Appendix A). Each draft was scored by at least two raters. Whenever the score difference between two raters was more than 1.5, discussion was held to resolve the disagreement. A mean score was used for analysis in the case of small (i.e., 1 point) discrepancies. Since the analysis focused on the seven high-level comment topics and these topics were only related to four of the dimensions, the scores on these four dimensions were used to represent assessor dimensional writing performance: Thesis (*Kappa* = 0.79), Evidence (*Kappa* = 0.76), Organization (*Kappa* = 0.71), and Explanation (*Kappa* = 0.69). The other two high-level dimensional scores, Argument and Rhetorical strategies, did not represent dimensional performance in the analyses.

Students' overall document scores were also calculated as the average of the full set of two low-level dimensional scores and six high-level dimensional scores (*Kappa* = 0.67). The overall document scores were used to represent students' genre writing performance.

3.4 Procedure

Following best instructional practices, training on effective peer review was conducted prior to the study with both teachers (by the research team) and then with students (by the teachers). The training workshop for teachers began with introducing the online peer assessment system, including a detailed description of the system and how to use the system to conduct peer review. All teachers acted as students during the workshop, uploading their own responses to a writing prompt, and then responded to peer (i.e., other teachers in the workshop) writing according to the rubrics provided by the system, which were the same peer review rubrics used by students.

Protocols (i.e., PowerPoint slides and scripts) were then given to instructors for how to train students on how to use the system and give effective feedback. In particular, students were instructed in class on how to use the reviewing system to turn in their first drafts and provide feedback to others. Good peer review practice was introduced, including a description of good quality feedback (i.e., specific and constructive) and poor feedback (i.e., vague and general). Then, students received an example student essay and feedback comments for that essay. They discussed as a whole class whether the comments could help revisions and what made the comments helpful. After discussion, students were required to provide feedback to a new essay based on the given rubrics. Then they discussed their feedback and ratings so that they could develop better understanding of effective peer review and the specific rubrics. The detailed rubrics were new to all the schools, and a common training process for the rubrics and the general reviewing approach was applied across schools just prior to the study. All the students therefore had a similar amount of experience with receiving and providing feedback based on the given rubrics.

3.5 Data analysis

RQ1 analyses. To answer RQ1 (identify empirically separable grain sizes of assessor writing performance), a correlation analysis on assessor topic performance was conducted.

RQ2 analyses. To answer RQ2 (i.e., the relationships of assessor writing performance at different grain sizes with problem identification accuracy), 234 assessors' data (the feedback they provided, their performance on each of the seven comment topics, and document topic quality) were analyzed to calculate hit and false alarm rates for each topic according to Signal Detection Theory (SDT, Green & Swets, 1966). When responding to essays, identifying a problem is not synonymous with identifying the problem correctly, especially for novice writers. SDT is a very commonly-used framework to analyze identification accuracy. SDT was first developed and applied in studies of perception, and SDT has grown to be used in many areas to examine observers' decision-making behavior wherever there is ambiguity and potential for bias in identification and recognition tasks (e.g., recognition memory, Higham et al., 2009). An observer is conceptualized as needing to answer "yes" or "no" to the presence of the signal against a background of noise. Correctly identifying a signal when it is actually present is called a *hit (H)*, and not identifying a signal when it is actually present is called a *miss (M)*. Correctly indicating that a signal is not present when it is actually absent is called a *correct rejection (CR)*, and indicating that a signal is present when it is actually absent is called a *false alarm (FA)*. Hit rates are then determined by dividing the total number of hits by the total number of hits and misses, i.e., $H/(H+M)$. False alarm rates were determined by dividing the total number of false alarms by the total number of false alarms and correct rejections, i.e., $FA/(FA+CR)$.

To fit the present/absent categories required by SDT, the three performance categories (i.e., Competent, Ok, Poor) within each of the 7 comment topics were recategorized into two categories: Ok and Poor categories were combined into a Less Competent category in statistical analyses because both categories indicated that the essays needed to be improved on the given topic. In addition, Ok and Poor performance groups were found to show similar patterns in terms of the relationship between problem identification and feedback quality.

To visually explore patterns of assessor performance on problem identification accuracy in each of the seven topics, a hit rate and false alarm rate was calculated for each overall assessor performance group on each topic. Assessor performance level was defined by performance on the specific topic by the assessor. For example, was the hit rate for identifying problems in the "Quality of explanations" higher for student assessors whose own explanations in their essays were generally of high quality than students whose own explanations were not generally of high quality?

Since there is nested structure to the student data (i.e., students are nested within schools), two-level logistic regressions were conducted with problem identification (i.e., hits vs. misses) as the binary outcome. The variables were divided into student-level variables (i.e., Level 1) and school-level variable (i.e., Level 2). Level 1 variables included assessor topic performance, assessor dimensional performance, and document dimensional quality. The main predictors of assessor writing performance were assessor topic performance and assessor dimensional performance, while document dimensional and overall quality were considered as potential control variables. However, as document dimensional quality and document overall quality were strongly correlated (r = 0.75; see Table D1 in Appendix D), document overall quality was dropped from the logistic regression. Document topic quality was not included because an assessor only needed to identify a problem in a comment when the assessee was not fully competent on the comment topic. Dummy codes for dimension (i.e., Explanation, Organization, Evidence, Thesis) were included as level 1 control variables so that the potential impact of the confounding variables could be removed (e.g., some dimensions may be less likely to be identified correctly and coincidentally also more likely to occur in the documents). School context (i.e., whether the school is title I school) was a level 2 variable. As a robustness check, the regression model was conducted twice: initially without and then with comment dimension and school context variables. Effect sizes in these logistical regression models are presented in terms of odds ratios (how the odds of having a 1 vs. 0 in the binary outcome variable vary with each unit of increase in the predictor variable). An odds ratio of 1 indicates no effect, an odds ratio less than 1 indicates a negative relationship, and an odds ratio greater than 1 indicates a positive relationship.

RQ3 analyses. It is important to acknowledge the complex nested nature of this data: comments nested inside of assessors nested inside of schools. To answer RQ3 (i.e., the relationship of assessor writing performance at different grain sizes with helpfulness of feedback), three-level logistic regression models were conducted on the feedback helpfulness binary outcome variable. The examined variables were divided into comment-level variables (i.e., level 1), student-level variables (i.e., level 2), and school-level variable (i.e., level 3). Similar to the RQ2 analyses, the main potential predictor variables included assessor topic performance, assessor dimensional performance, and assessor genre performance. Document dimensional and overall quality entered the model as control variables. We also included the interaction of assessor dimensional performance and document dimensional quality since it may be difficult to provide useful feedback to a document with higher dimensional quality. As in the previous analysis, document topic quality was not included in the models because no problem on a particular topic needed to be identified if the assessed document did not have a problem with the given topic.

Strong correlations were found between assessor dimensional performance and assessor genre performance (*r* = 0.78), and between document dimensional quality and document overall quality (*r* = 0.75) (see Table D1 in Appendix D). To avoid multicollinearity problems, assessor genre performance and document overall quality were dropped (as redundant with the corresponding dimensional competences) from the logistic regression models. All continuous variables (see Table 1) were grand mean centered for analysis.

As a robustness check, the regression model was conducted twice: initially without and then with comment-level and school-level control variables. Similar to RQ2, school context was included as a control variable. However, different from the RQ2 analysis because RQ3 was conducted at the level of specific comments rather than comment topics, additional comment-level control variables included the five coded comment features, comment length, and the seven comment topics as dummy codes. Teasing out the potential influence of the control variables can help us better understand the specific impact of assessor writing performance on feedback helpfulness.

A summary of the main statistical analyses used to address each of the three research questions is presented in Table 2. Appendix E presents a visualization of the data structure to show how problem identification accuracy and feedback helpfulness did not occur in the same subsets of data and thus could not be examined within one larger model or use similar control variables.

Insert Table 2

3.6 Transparency and Openness

In the prior section, we reported how we determined our sample size, data exclusions, manipulations, and measures, following JARS (Kazak, 2018). The study was reviewed and approved by as exempt under the educational exemption by Human Research Protection Office at the University of Pittsburgh. Teachers were paid for their participation in the study. The data and analysis code are available at https://osf.io/ta3cu/. Data were analyzed using Stata 16. This study's design and its analysis were not pre-registered.

4. Results

4.1. RQ1: Separable Grain Sizes of Assessor Writing Performance

To answer the first research question, a correlation analysis was conducted to investigate the correlations in performance on different problem topics (i.e., assessor writing performance on the focal seven feedback topics; see Table 3). This correlation analysis showed that performance on the two thesis-related topics (i.e., thesis clarity and connection in thesis) were strongly correlated (*r* = .71) (see Table 3). Moderate relationships were observed among assessor performance on other topics (e.g., quantity of explanation and quality of explanation, quantity of evidence and quantity of explanation). Overall, performance on different topics were strongly or moderately related to one another when they were on conceptually connected topics. For example, if the thesis is not clear (i.e., topic: thesis clarity), it will be hard for the author to make a specific or clear connection between rhetorical strategies and the author's argument (i.e., topic: connection in thesis; see Appendix B). Similarly, quantity of evidence and quality of explanation (i.e., explanation of evidence) were correlated (*r* = .57) primarily because a higher quality explanation depends upon having a certain amount of evidence to explain. Despite the interrelationship of assessor performance on the topics, the performance within each topic conceptually focuses on different aspects. For example, thesis clarity focuses on meaning expression, while connection in thesis focuses on the two key components (i.e., rhetorical strategies that the author of the source passage used and the main argument of the source passage) involved in the thesis. Therefore, due to the targeted foci and potential differential effects of performance within each specific topic, we treated the seven topic performance variables as sufficiently separable for the purposes of RQ2 and RQ3.

Insert Table 3

4.2 RQ2: The Relationships of Assessor Writing Performance at Different Grain Sizes with Problem Identification Accuracy

Two analyses were conducted to answer the second research question (see Table 2). First, assessors' hit and false alarm rates for each topic were calculated, which involved data from 234 assessors. Second, two-level logistic regression analysis models were conducted with problem identification accuracy as the outcome and assessor topic and dimensional performance as the main predictors. Regression Model 1 tested the relationship of assessor topic and dimensional performance to problem identification accuracy when controlling for document dimensional quality. As a robustness check, Regression Model 2 tested the relationship of assessor topic and dimensional performance to problem identification accuracy when controlling for more potential confounds with problem identification accuracy: document dimensional quality, dimensions, and school context.

According to the topic-specific analyses of hits rates, lower performing assessors significantly identified more (not fewer) problems on three topics: for Connection in thesis, $\chi^2(1) = 4.11$, $p < .05$; for Body connected to thesis, $\chi^2(1) = 6.16$, $p < .05$; for Transitions, $\chi^2(1) = 9.12$, $p < .01$ (see Figure 4A). Lower performing assessors also had directionally higher false alarm rates on three topics than did higher performing assessors (Connection in thesis, Body connected to thesis, Transitions), but those effects were not statistically significant (see Figure 4B). Thus, assessor topic performance did not appear to generally produce higher accuracy in identification of problems. Rather, lower performing students appeared to bring up some topics more often, whether appropriate to do so or not. Note that mean performance for each of the topics and dimensions (when focusing on just the assessors who choose to focus on a given topic) was roughly the same, with means just below 5 on each of the 1–7 rubric dimensions and means around 1 on the Poor=0 to Competent=2 topic performance scale (see Table D1 in Appendix D). Thus, the variation in hit rates by topic was not a simple function of general performance levels on each topic, again supporting the notion that topic performance was not the driver of accuracy of problem identification.

Insert Figure 4A and Figure 4B

Correlations among assessor writing performance, document quality, and feedback dimensions were examined to reveal potential confounds and multicollinearity problems among the predictor variables. According to the correlations in the review-level predictors (see Table D1 in Appendix D), assessor topic performance was moderately correlated with assessor dimensional performance (*r* = .33) and assessor genre performance (*r* = .26), while assessor genre performance and assessor dimensional performance was strongly correlated (*r* = .78). The results implied that assessor writing performance on specific topics was comparatively independent from their dimensional and genre writing performance. No strong correlations were found between assessor writing performance or document quality and comment dimensions. Assessor genre performance and document overall quality were not included in the logistic regressions to avoid multicollinearity problems because of the strong correlations they had with assessor dimensional performance and document dimensional quality.

Two-level logistic regression models were conducted to further explore whether assessor topic performance or assessor dimensional performance predicted problem identification accuracy. Inspection of the included predictor variables indicated no significant multicollinearity problems.

Logistic regression analyses indicated that neither assessor topic performance nor document dimensional quality were related to problem identification accuracy (see Table 4). Further, even assessor dimensional performance was not a significant predictor when control variables (i.e., dimensions and school) were included. Thus, as suggested by the visualization in Figures 4A and 4B, assessor performance (at any grain size) did not seem to be significantly related to identification accuracy. At most, some dimensions (explanation and evidence) involved easier-to-accurately-identify problems.

Insert Table 4

RQ3: The Relationships of Assessor Writing Performance at Different Grain Sizes with Helpfulness of Feedback

Three-level logistic regression models were conducted to answer the third question with feedback helpfulness as the outcome, and assessor topic and dimensional performance as the key predictors (see Table 2). Model 1 tested the relationship of assessor writing performance to feedback helpfulness while controlling for document dimensional quality. Model 2 tested the relationship of assessor writing performance to feedback helpfulness while controlling for document dimensional quality, feedback features, feedback topics, and school context.

Correlations among the variables operationalized at the comment-level (see Table D2 in Appendix D) showed that feedback helpfulness was significantly correlated with assessor genre performance, assessor dimensional performance, and assessor topic performance, in addition to being significantly correlated with general comment length and with having the comment features of including an explanation or a suggestion. Interestingly, document dimensional quality was negatively correlated with feedback helpfulness, although it was a very small correlation (r = - .06). Because of the strong correlation between assessor dimensional performance and assessor genre performance (r = .80) as well as between document dimensional quality and document overall quality (r = .81), only assessor topic performance, assessor dimensional performance, and document dimensional quality were included in the regression models to avoid multi-collinearity problems. The correlations between assessor writing performance/document quality and feedback features were small (r < .2). Associated with the feedback features indicator variables, there were two moderate correlations: identification with explanation (r = .49), and explanation with comment length (r = .42).

The logistic regression analyses formally testing RQ3 revealed that assessor topic performance, assessor dimensional performance, and document dimensional quality were significant factors in predicting feedback helpfulness (see Table 5). Assessor topic performance was a stronger predictor than assessor dimensional performance, as predicted. Document dimensional quality was negatively correlated with feedback helpfulness: the weaker the document was on a dimension, the more likely the author was to receive helpful feedback. However, the interaction of assessor performance and document quality was not significant. To visualize the effect sizes of the relationship of assessor performance at different grain sizes to feedback helpfulness and directly compare those effect sizes to the ones observed for predicting problem detection accuracy, Figure 5 graphs all the relevant odds-ratios from the different logistic regressions. Assessor dimensional performance was associated with a 25% increase in the likelihood of helpful feedback, but assessor topic performance made helpful feedback more than twice as likely.

Insert Table 5 and Figure 5

In terms of control variables, feedback helpfulness was significantly related to whether the comment included an explanation, suggestion, solution (negatively), and comments from students at Title I school were notably higher. The positive correlations of the presence of an explanation, a suggestion, and being at a Title I school with feedback helpfulness are consistent with prior research investigating feedback features and feedback quality (e.g., Gao et al., 2019; Gielen & De Wever, 2015a; Prins et al., 2006; Wu & Schunn, 2020c). Most importantly, the predictive relationship of the performance variables was robust across models that included these highly significant correlates of comment usefulness.

5. General Discussion

Peer review has been widely and increasingly used as an efficient means for helping students improve writing proficiency, but often instructors and students have doubts about whether peer assessors as novice writers are able to correctly identify problems or provide helpful feedback. Understanding the relationship of assessors' writing performance at different grain sizes to the problem identification accuracy and helpfulness of their peer feedback is important in guiding instructors' effective use of and students' active engagement in peer feedback. Figure 6 summarizes the findings regarding significant predictors of identification accuracy. Overall, the results indicated that assessor writing performance (at any grain size) did not predict problem identification accuracy, but higher performing assessors at the level of dimensions and specific topics were more likely to provide useful feedback. Thus, this study both replicates prior work showing that general writing performance shows little effect, but also validates concerns of instructors and students as having a clear (but partial) basis in reality.

Insert Figure 6

5.1 Separable Grain Sizes of Assessor Writing Performance

For peer assessment research, the current study provides a methodology for uncovering appropriate grain-sizes of writing performance within a context and shows that, in at least this context, the effects of two grain sizes of writing performance could be empirically distinguished, separating the finest grain size of specific problem topics from even the slightly more aggregate dimensional performance. Prior research on assessor writing performance generally focused only on assessors' overall writing performance without consideration of other grain sizes (e.g., Allen & Mills, 2014; Dijks et al., 2018; Wu & Schunn, 2020a, c). The current study draws attention to the importance of considering much more fine-grain sizes of assessor writing performance: dimensions and topics.

It should be noted that the seven most-common comment topics examined in the study were closely aligned with the peer review rubrics and that most of these rubrics were closely aligned with the writing assignment's genre. It may be that the specific problem topic may be less separable from overall writing performance when more general writing rubrics are used to guide the peer reviewing or when the focus is on lower-level aspects of writing (i.e., spelling and grammar) or when there is wider variation in overall writing performance (e.g., in Massive Open Online Course). Genre specific rubrics draw assessors' attention to the most salient problems of a particular writing task and the specific context of peer feedback (Yu, 2021).

5.2 Assessor Performance and Problem Identification Accuracy

The finding regarding assessor writing performance and identification accuracy was not consistent with the initial predictions. That is, neither assessor dimensional writing performance nor assessor topic-specific writing performance predicted problem identification accuracy. One explanation for the finding might be that students, regardless of their own writing performance, know what makes a good piece of writing when given the evaluation supports provided by the reviewing rubric. Descriptions of writing process models provide some support for this possibility: "Evaluation is also a constructive, goal-driven process guided by the intentions (e.g., goals and criteria) one brings to the task" (Flower et al., 1986: 36). It might be that peer review training together with well-structured criteria provided the assessors with sufficiently clear goals so that they could identify problems and thereby improved feedback validity. Further, the analytic focus in this study on the most-commonly-mentioned comment topics may have resulted in attending to issues that were generally within the students' zone of proximal development (Lundstrom & Baker, 2009)—knowledge and topics that students are on the verge of mastering. In support of this idea, some other research suggests that students rarely comment upon the kinds of problems that are a challenge to most students in the class (Gao et al., 2019). Therefore, instructors could wisely choose to focus their own comments on the especially pervasive issues across papers, since those issues will likely be less commonly addressed by student comments.

The lack of significant relationships of assessor writing performance and identification accuracy further supports the idea that writing performance is not equal to reviewing performance (Wu & Schunn, 2019). As noted by Sadler (1989), "There are many domains of human activity where people are expert at appraising existing objects, sometimes in a highly sophisticated way, but are themselves incapable of producing objects of the type in question" (p. 139). That reviewing performance is higher than writing performance is also consistent with Yu's (2021) study, which revealed that students' writing performance problems and lack of knowledge of academic writing (genre texts) did not prohibit them from judging peers' work because students could draw on other sources for help (e.g., rubrics, feedback from other peers, reference books). Since students often failed to identify problems in the documents, future research could collect additional qualitative data such as think-aloud protocols or stimulated recall interviews to explore what other factors might be limiting assessors' problem identification accuracy, such as reviewing strategies, time constraints, and available external supports.

5.3 Assessor Performance and Helpfulness of Feedback

Both assessor dimensional and topic writing performance were found to be significant predictors of helpfulness of feedback, with a stronger relationship with specific topic writing performance, as initially predicted (see Figure 6). When an assessor did not have at a particular writing problem (e.g., problems with Quality of explanation or Quantity of explanation) or on a dimension (e.g., Explanation), they were more likely to provide helpful feedback on that problem topic. The findings were consistent with our predictions and some prior research regarding quality of feedback (e.g., Yu, 2021) and perceived quality of feedback (e.g., Dijks et al., 2018). Higher performing assessors likely have a better understanding of criteria and how a particular problem area could be successfully addressed (Yu, 2021). That the effect of assessor performance on feedback helpfulness is much stronger at the specific problem grainsize suggests the effect is likely driven by knowledge of revision strategies rather than just understanding of criteria, since the dimensions align with the criteria.

The finding may also be explained by assessors' confidence in their writing ability, which can influence the type of feedback they provide (Allen & Katayama, 2016). Previous research in ESL writing (e.g., Allen & Mills, 2014) and English native speakers' writing (e.g., Patchan & Schunn, 2015) has found that lower performing assessors provide fewer suggestions on peer writing than higher performing assessors, especially when higher performing assessors are paired with lower-performing assessors. In other words, the feedback of lower performing assessors might have been less helpful because they were not sufficiently confident to suggest a revision or carefully explain the problem rather than not having useful suggestions or explanations they could have provided. If future research supports such an explanation, additional assistance could be provided to help lower-performing assessors raise their confidence and thereby provide more helpful feedback.

The metaphorical phrase that students and instructors often use when they express doubt about the usefulness of peer feedback is “the blind leading the blind”. The current findings place in sharp relief the ways in which this metaphor is inaccurate for peer feedback. Students at varying performance levels are just as likely to notice problems, so their ability to *see* problems, particularly with carefully designed rubrics, is not the issue. Rather, the issue is resolving the problem. So, perhaps a more apt metaphor would be one of receiving driving advice from a non-driver.

Now that the empirical phenomenon is established, there remain open theoretical questions about what caused the effect. Is it that peers do not have a useful solution to suggest, or do they struggle in clearly articulating a useful solution? The unexpected finding that solutions were negatively correlated with feedback helpfulness is supported by the work of Tseng and Tsai (2007), who observed that receiving solutions seemed to be harmful for high school students’ subsequent learning. The research finding could also be explained by students’ self-efficacy. For example, Wang and Wu (2008) found that self-efficacy predicted students’ use of learning strategies and elaborated peer feedback behavior. Alternatively, are peers less likely to offer a solution because they lack confidence? Allen and Katayama (2016) found that lower performing students are not confident of their ability and less willing to participate in peer feedback. Further research is required to unpack the underlying mechanisms.

6. Implications and Conclusions

6.1 Implications for Practice

Peer review plays an important role in writing instruction, but there remain common suspicions towards whether students as novice writers are sufficiently high performing to identify problems in peer' work and provide useful feedback to improve document quality. The influence of assessor writing performance at different grain sizes on problem identification accuracy and helpfulness of feedback has been ignored in empirical research. The current study established ways of distinguishing assessor writing performance at three grain sizes (i.e., genre, dimension, and topic), finding that an assessor with strong overall genre writing performance is not necessarily strong in all dimensions or in all specific problems within a dimension. Further the study showed that assessor dimensional performance and topic performance were not correlated with problem identification accuracy, but predicted feedback helpfulness. The finding that both higher and lower performing assessors were equally able to identify problems correctly should be communicated to both students and instructors to motivate them to engage with peer feedback. That is, there appears to be no need to worry about involving lower performing assessors in peer feedback when well-structured peer feedback training and reader-friendly rubrics are used. Although lower performing assessors might be less likely to provide useful feedback, the amount of incorrect feedback was small (approximately 2% of the comments in the present study). In addition, it is worth remembering other useful aspects of peer feedback: "imperfect feedback from a fellow student provided almost immediately may have much more impact than more perfect feedback from a tutor four weeks later" (Gibbs & Simpson, 2004: 19).

Second, the current findings suggest that teachers should conduct writing instruction and peer review training in terms of rhetorical knowledge of specific genres and specific topics of a genre assignment. Positive effects of peer response with instruction in specific genre knowledge on better text quality have been found in prior research (e.g., Hoogeveen & van Gelderen, 2015).

Instruction in specific genre knowledge can improve students' genre awareness and help students focus on concrete problems when they are writing and commenting on peers' work (Hoogeveen & van Gelderen, 2015; Yu, 2021). In addition, providing focal training on specific topics would both decrease their frequency in a subsequent assignment as well as increase the helpfulness of peer feedback on that subsequent assignment. Genre-focused rubrics on specific problems, especially those high-frequency problems, should also be emphasized in peer feedback training. Third, the current study suggests that explanations and general suggestions should also be emphasized in peer review training, and cautions need to be taken when very specific solutions are provided. For example, instructors can incorporate guiding questions (e.g., why do you think it is a problem?) in peer review to help assessors include more information in peer feedback (Gielen & De Wever, 2015b). Further, how to identify and explain problems of an unsuccessful document could be demonstrated in the context of the rubrics (Min, 2016).

Fourth, it may be productive to assign assessors to assessees by taking into account their dimension or topic writing performance. Ohta (2001) suggested that, "No learner is universally more or less capable than a peer, but that each learner presents an array of strengths and weaknesses that may be complementary" (p. 76). The current research suggests it is non-productive to categorize students as higher and lower performing learners in terms of general language or writing performance (Watanabe, 2008). Additional functions of online peer review systems could be developed to automatically group assessors and assessees taking into account their strengths and weaknesses on specific genre problems (Wu & Schunn, 2021). In addition, to deal with issues of low trust or lack of confidence in peer assessors, students could be encouraged to provide feedback based upon their strengths, which would be more likely to be helpful feedback.

Lastly, instructors could highlight the contributions of both higher and lower performing students and raise their confidence regarding the benefits of peer feedback to enable them to leverage the learning activity. Lower performing students' lack of confidence can result in less engagement with peer feedback and reluctance to provide feedback (Allen & Katayama, 2016). Negative perceptions of peers' writing ability may not only limit students' participation in peer feedback activities but may also deprive them of the opportunity to exploit its affordances (Dijks et al., 2018). Teachers should encourage students to hold positive perceptions of peer feedback and take the activity seriously by sharing the benefits of peer feedback for both higher and lower performing students.

6.2 Limitation

There are several limitations to be acknowledged in the current study. First, different assessors reviewed different essays rather than all being given the same piece of writing in the present study. The essays could have such widely varying qualities such that different treatable issues could result in different problem identification accuracy and feedback helpfulness, and the current regression approach may not have fully controlled for those differences. Future lab-based research could more directly control the quality of the drafts so that the amount and types of problems that can be identified are entirely identical. For example, assessors could all be required to comment on a common set of documents (e.g., Wu, 2019).

Second, the present study was conducted in a reciprocal context in which students' roles of assessors and assessees were interactive. Assessor competence in identifying problems and providing helpful feedback are partially the result of interactive effects of receiving and providing feedback simultaneously rather than assessors have an unchanging writing performance (Huisman et al., 2017). Future research could compare assessors' problem identification accuracy and feedback helpfulness before and after they receive feedback.

Third, assessor writing performance is not equal to assessor reviewing performance. Although identical elements have been identified conceptually between writing and reviewing (Patchan & Schunn, 2015), an assessor may be a strong writer but still be less effective at providing feedback due to motivation or lack of revision strategies. It is also possible that learners are good at some dimensions but fail to identify some problems because of their attention is on other salient problems. Future research should seek to further distinguish assessor writing performance from reviewing performance.

Fourth, we also note that the present study only focused on high frequency problems and this focus was the largest reduction in the subset of data that was analyzed. The findings observed here may not generalize to lower frequency problems that collectively still accounted for many comments. However, it is important to note that the frequency of the seven problem topics did vary substantially and yet the findings were relatively robust across the seven different problem topics. On a related point, no comments about low-level writing issues were included in the analyses because of the limited number of comments on any given low-level problem. Future research could collect more data on low-level comments and investigate assessors' identification accuracy and feedback helpfulness. Similarly, future research could also examine the effects of assessor writing performance on other aspects of review quality besides identification accuracy and comment helpfulness, such as relative focus on higher vs. lower level writing problems, comment implementability, amount of feedback, and accuracy of ratings.

Finally, the writing task analyzed in the present study was a source-based rhetorical strategy analysis, which requires students to analyze rhetorical strategies used by the author of the source passage. Therefore, the content of the writing task and the specific problems included in peer feedback were all related to a provided source passage and students' knowledge of rhetorical strategies. The research findings could be different if the writing task was less technical or relied more heavily on outside knowledge. For example, the content (e.g., thesis, evidence) of scientific writing or argumentative writing often does not rely on any provided source passage. With such writing tasks, students present a thesis and evidence related to the topic, and provide feedback primarily based on their own understanding and general knowledge of the topic. Thus, more general knowledge may be more important in generating accurate or helpful feedback in such a writing task.

6.3 Conclusion

The present study investigated the relationship of assessor writing performance at three grain sizes (i.e., genre performance, dimensional performance, topic performance) to problem identification accuracy and helpfulness of peer feedback. Higher and lower performing assessors did not show significant differences in problem identification accuracy. In terms of feedback helpfulness, assessor topic performance and assessor dimensional performance were significant predictors. Based on the findings, instructors are encouraged to continue to include students with low writing performance as peer feedback providers. Genre-based rubrics should be used to help students better understand rhetorical knowledge of different genres and provide genre specific feedback that is aligned with the writing task. In addition, assessors could be assigned to assessees based upon their specific topic writing performance rather than general writing performance to maximize the benefits of peer feedback.

References

Allen, D., & Katayama, A. (2016). Relative second language proficiency and the giving and receiving of written peer feedback. *System*, 56, 96–106. http://dx.doi.org/10.1016/j.system.2015.12.002

Allen, D., & Mills, A. (2014). The impact of second language proficiency in dyadic peer feedback. *Language Teaching Research*, 4, 1–16. http://dx.doi.org/10.1177/1362168814561902

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

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

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. (2020). Program summary report. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/research/2020/Program-Summary-Report-2020

Dijks, M. A., Brummer, L., & Kostons, D. (2018). The anonymous reviewer: The relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assessment & Evaluation in Higher Education*, 43(8), 1258-1271. https://doi.org/10.1080/02602938.2018.1447645

DiPardo, A., & Freedman, S. W. (1988). Peer response groups in the writing classroom: Theoretic foundations and new directions. *Review of Educational Research*, 58, 119–149. https://doi.org/10.3102/00346543058002119

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review,* 32, 481–509. https://doi.org/10.1007/s10648-019-09510-3

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, 37(1), 16–55. https://doi.org/10.2307/357381

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

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, 1, 3–31. http://eprints.glos.ac.uk/id/eprint/3609

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

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

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology*, 35(1), 57–80. https://doi.org/10.1037/a0013865

Hoogeveen, M., & Van Gelderen A. (2015). Effects of peer response using genre knowledge on writing quality: A randomized control trial. *The Elementary School Journal*, 116(2), 265–290. https://doi.org/10.1086/684129

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

Huisman, B., Saab, N., Van Driel, J., & Van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development*, 36(7), 1433–1447. https://doi.org/10.1080/07294360.2017.1325854

Huisman, B., Saab, N., Van den Broek, P., & Van Driel, J. (2018). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education*, 44, 863–880. https://doi.org/10.1080/02602938.2018.1545896

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

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1–2. http://dx.doi.org/10.1037/amp0000263

Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., & Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing*, 48. https://doi.org/10.1016/j.jslw.2019.100700

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

Li, H., Xiong, Y., Charles, V., Guo, X., Lyu, Y., Rurik, T. (2020). Does peer assessment promote student learning: A meta-analysis. *Assessment & Evaluation in Higher Education*, 45(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

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

Liu, N., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290. https://doi.org/10.1080/13562510600680582

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

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, 18(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Min, H-T. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing*, 31, 43–57. http://dx.doi.org/10.1016/j.jslw.2016.01.004

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

Ohta, A. S. (2001). *Second language acquisition processes in the classroom: Learning Japanese*. Mahwah, NJ: Lawrence Erlbaum.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Education Evaluation*, 39(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science*, 41(2), 381–405. https://doi.org/10.1007/s11251-012-9236-3

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, 43(5), 591–614. https://doi.org/10.1007/s11251-015-9353-x

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research*, 8(2), 227–265. https://doi.org/10.17239/jowr-2016.08.02.03

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

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1

Schunn, C. D. (2016). Writing to learn and learning to write through SWoRD. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive Educational Technologies for Literacy Instruction*. NY: Taylor & Francis, Routledge.

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

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–89. https://doi.org/10.3102/0034654307313795

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(4), 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

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *High Education*, 76, 467–481. https://doi.org/10.1007/s10734-017-0220-3

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 Bergh, Huub. (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, 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

Villamil, O. S., & De Guerrero, M. C. M. (1996). Peer revision in the L2 classroom: Social-cognitive activities, mediating strategies, and aspects of social behavior. *Journal of Second Language Writing*, 5(1), 51–75. https://doi.org/10.1016/S1060-3743(96)90015-6

Wang, S. L., & Wu, P. Y. (2008). The role of feedback and self-efficacy on web-based learning: The social cognitive perspective. *Computer & Education*, 51(4), 1589–1598. https://doi.org/10.1016/j.compedu.2008.03.004

Watanabe, Y. (2008). Peer-peer interaction between L2 learners of different proficiency levels: Their interactions and reflections. *The Canadian Modern Language Review*, 64(4), 605–635. https://doi.org/10.1353/cml.0.0008

Wooley, R. S., Was, C. A., Schunn, C. D., & Dalton, D. W. (2008). *The effects of feedback elaboration on the giver of feedback.* Paper presented at the 30th Annual Meeting of the Cognitive Science Society, Washington DC.

Wu, Y., & Schunn, C. D. (2019). The learning science of multi-peer feedback for EFL students. *Technology Enhanced Foreign Language Education*, 189, 13–21.

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

Wu, Y., & Schunn, C. D. (2020b). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal*, 58, 492–526. https://doi.org/10.3102/0002831220945266

Wu, Y., & Schunn, C. D. (2020c). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology*, 62. https://doi.org/10.1016/j.cedpsych.2020.101897

Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science*, 49, 365–394. https://doi.org/10.1007/s11251-021-09546-5

Wu, Z. (2019). Lower English proficiency means poorer feedback performance? A mixed-methods study. *Assessing Writing*, 41, 14–24. https://doi.org/10.1016/j.asw.2019.05.001

Yu, S. (2021). Giving genre-based peer feedback in academic writing: Sources of knowledge and skills, difficulties and challenges. *Assessment & Evaluation in Higher Education*, 46(1), 36–53. https://doi.org/10.1080/02602938.2020.1742872

Yu, S. & Lee, I. (2016). Understanding the role of learners with low English language proficiency in peer feedback of second language writing. *TESOL Quarterly*, 50(2), 483–494. https://doi.org/10.1002/tesq.301

Zhao, H. (2011). Using learners' diaries to investigate the influence of students' English language proficiency on peer assessment. *Journal of Academic Writing*, 1(1), 126–134. https://livrepository.liverpool.ac.uk/id/eprint/2014347

Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education*, 45(3), 372–386. https://doi.org/10.1080/02602938.2019.1644603

Zhu, Q., & D. Carless. (2018). Dialogue within Peer Feedback Processes: Clarification and Negotiation of Meaning. *Higher Education Research & Development*, 37(4), 883–897. https://doi.org/10.1080/07294360.2018.1446417

Zong, Z., Schunn, C. D., & Wang, Y. (2021). Learning to improve the quality of peer feedback through experience with peer feedback. *Assessment & Evaluation in Higher Education*, 46(6), 973–992. https://doi.org/10.1080/02602938.2020.1833179

OSF. https://osf.io/ta3cu/

Table 1. Type, level, and definitions of outcome variables, key predictors, and control variables

Variable	Type	Level	Definition
Outcome variables			
Helpfulness of feedback	Binary	Comment	Whether a comment improved essay quality
Hit	Binary	Review	For essays having the given topic problem, 1 if the problem was identified, 0 if it was not identified
False alarm	Binary	Review	For essays not having the given topic problem, 1 if the problem was identified, 0 if it was not identified
Key predictors (Assessor writing performance)			
Assessor topic performance	Binary	Essay	Whether a student was competent on a specific topic
Assessor dimensional performance	Continuous	Essay	Score on a specific dimension
Assessor genre performance	Continuous	Essay	Score on an essay
Document quality control variables			
Document topic quality	Binary	Essay	Whether an assessed document has a problem with the given topic
Document dimensional quality	Continuous	Essay	The quality of the assessed document on the given dimension
Document overall quality	Continuous	Essay	Overall quality of an assessed document
Feedback feature control variables			
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
Comment length	Continuous	Comment	Words involved in a comment
School-level control variables			
School title	Binary	Assessor	Whether students came from Title I school

Table 2. Summary of research questions, data analysis methods, and included variables

Research questions	Data analysis methods	Included variables
RQ1: Grain sizes of assessor writing performance	**Method:** Correlation analysis of assessor topic performance	**Writing performance for the seven most frequent comment topics:** • Thesis clarity • Connection in thesis • Body connected to thesis • Transitions • Quantity of evidence • Quantity of explanation • Quality of explanation
RQ2: Relationships of assessor writing performance at different grain sizes and successful problem identification accuracy	**Method 1:** Visualization of hit and false alarm rate per topic by assessor topic performance	**Independent variables:** • Assessor topic performance **Dependent variable:** • Problem identification (hit rate & false alarm rate)
	Method 2: Two-level logistic regression	**Independent variables:** • Assessor topic performance • Assessor dimensional performance **Control variables:** • Document dimensional quality • Comment dimension • School **Dependent variable:** • Problem identification accuracy
RQ3: Relationships of assessor writing performance on different grain sizes and helpfulness of feedback	**Method:** Three-level logistic regression	**Independent variables:** • Assessor topic performance • Assessor dimensional performance **Control variables:** • Document dimensional quality • Comment feature • Comment topic • School **Dependent variable:** • Helpfulness of feedback

Table 3. Pearson correlations among assessor topic performance.

	Variables	1	2	3	4	5	6	7
1	Thesis clarity	1.00						
2	Connection in thesis	**0.71****	1.00					
3	Body connected to thesis	**0.49****	**0.42****	1.00				
4	Transitions	**0.24****	**0.19****	**0.38****	1.00			
5	Quantity of evidence	0.09	**0.02**	**0.32****	**0.34****	1.00		
6	Quantity of explanation	**0.21****	**0.11**	**0.48****	**0.32****	**0.57****	1.00	
7	Quality of explanation	**0.33****	**0.22****	**0.58****	**0.42****	**0.51****	**0.63****	1.00
	M	1.23	1.24	1.09	1.03	1.34	1.03	0.72
	SD	0.73	0.77	0.67	0.55	0.69	0.74	0.69

Notes. The correlation analysis was conducted with N = 234 assessors. $*p < .05$, $**p < .01$

Table 4. Two-level logistic regression estimates (B, standard error, and odds ratio) of assessor topic performance, assessor dimensional performance, and document dimensional quality in predicting rates of problem identification, without (Model 1) and with (Model 2) feedback dimension and school context control variables

Predictors of problem identification	*Model 1*		*Model 2*	
	B *(SE)*	*Odds ratio*	*B* *(SE)*	*Odds ratio*
Mean outcome in logits (r_{00})	-0.68 (0.17)	0.51***	-0.25 (0.18)	0.78
Key predictors				
Assessor topic performance	-0.08 (0.09)	0.92	0.13 (0.10)	1.14
Assessor dimensional performance	0.09 (0.04)	1.09*	0.04 (0.04)	1.04
Document quality control variable				
Document dimensional quality	-0.01 (0.04)	0.99	-0.04 (0.04)	0.96
Comment Dimension Control Variables (reference: Evidence Dimension)				
Explanation Dimension	-	-	0.64 (0.12)	1.90***
Organization Dimension	-	-	-1.21 (0.13)	0.30***
Thesis Dimension	-	-	-0.73 (0.13)	0.48***
School-level Control Variable				
Title I school (Title I=1)	-	-	-0.37 (0.20)	0.69
Model fit statistics				
Log Likelihood	-2160.97		-1948.76	
AIC	4331.94		3915.53	

Note. The regression analyses were conducted with N = 234 assessors. - = variables not included in the model.
* *p*<.05, *** *p* < .001

Table 5. Three-level logistic regression analysis of assessor topic performance, assessor dimensional performance, and document dimensional quality in predicting helpful feedback, without (Model 1) and with (Model 2) feedback feature and school control variables

Predictors	Model 1		Model 2 (Controls included)	
	B (SE)	*Odds ratio*	*B (SE)*	*Odds ratio*
Mean outcome in logits (r_{00})	1.82 (0.20)	6.17***	2.25 (0.55)	9.51
Key predictors				
Assessor topic performance	0.94 (0.22)	2.55***	0.88 (0.23)	2.41***
Assessor dimensional performance	0.25 (0.08)	1.29**	0.22 (0.09)	1.25**
Document quality control variables				
Document dimensional quality	-0.18 (0.07)	0.83*	-0.17 (0.08)	0.84*
Assessor dimensional performance * Document dimensional quality	-0.02 (0.07)	0.98	-0.02 (0.07)	0.98
Comment-level control variables				
Feedback features				
Mitigating praise	-	-	0.04 (0.15)	1.04
Identification	-	-	0.14 (0.19)	1.15
Explanation	-	-	0.50 (0.18)	1.65**
Solution	-	-	-0.71 (0.26)	0.49**
Suggestion	-	-	0.56 (0.17)	1.76**
Comment length	-	-	0.00 (0.00)	1.00
Feedback topics				
Quantity of evidence	-	-	-1.29 (0.52)	0.27*
Quantity of explanation	-	-	-1.35 (0.51)	0.26**
Quality of explanation	-	-	-1.23 (0.50)	0.29*
Body connected to thesis	-	-	-0.75 (0.58)	0.47
Transitions	-	-	-1.33 (0.58)	0.26*
Thesis clarity (reference: Connection in thesis)	-	-	-1.19 (0.55)	0.30*
School-level control variables				
Title I school (Title I=1)	-	-	0.48 (0.22)	1.62*
Model fit statistics				
Log Likelihood	-776.56		-752.03	
AIC	1567.13		1544.05	

Note. Logistic regression analyses were conducted with N = 1,921 comments. - = variables not included in the model. * p < .05, ** p<.01, *** p < .001

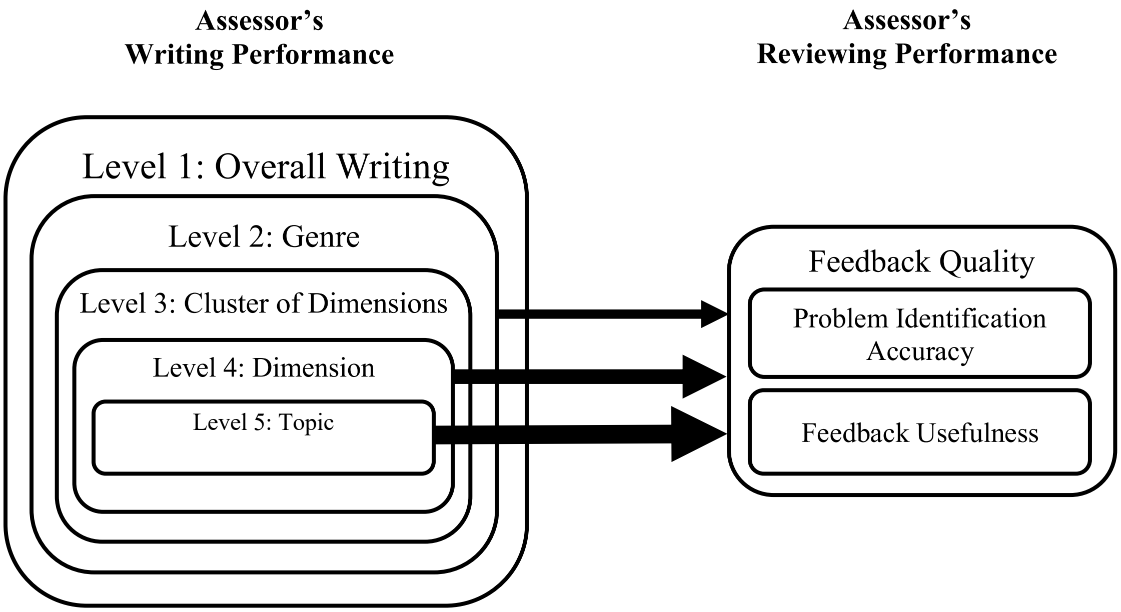


Fig 1. Assessor writing performance at different grain sizes (from macro=large font to micro=small font) and their potential relationship to feedback quality. Line thickness of the arrows connecting assessor writing performance with feedback quality corresponds with expected strength of relationship.

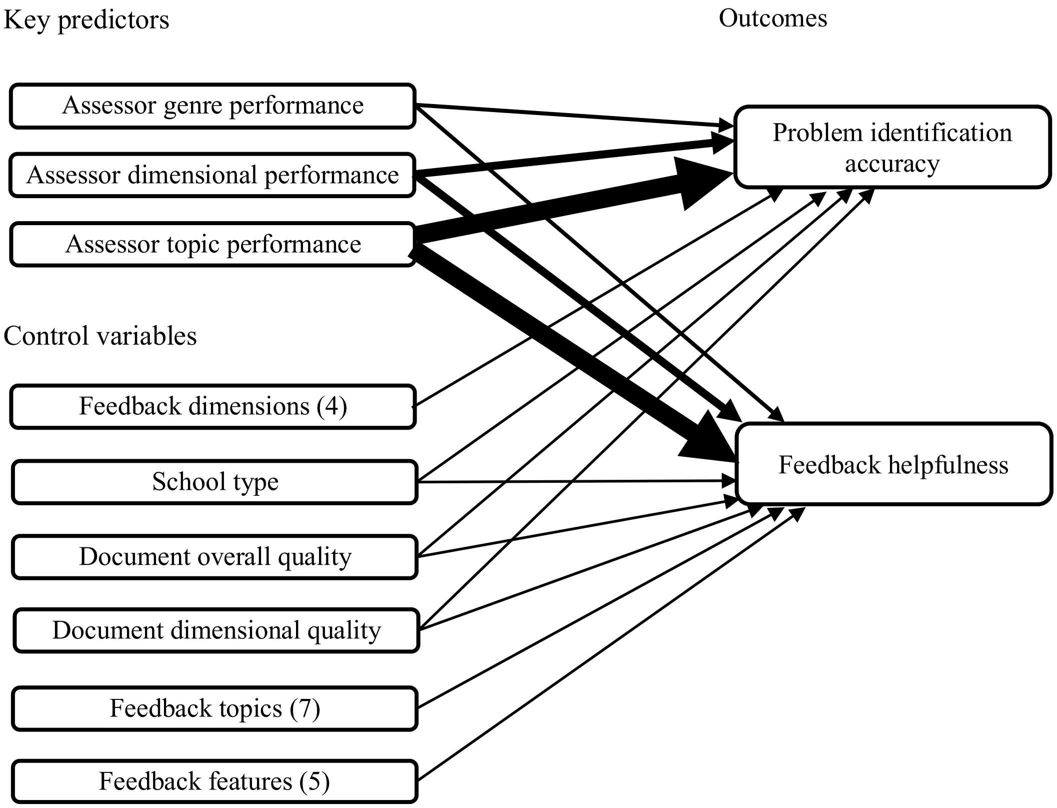


Fig 2. Tested model of the relationships between assessor writing performance at three different grain sizes with problem identification accuracy and usefulness of peer feedback. There are four feedback dimensions, seven feedback topics, and five feedback features. Line thickness corresponds with expected strength of relationship.

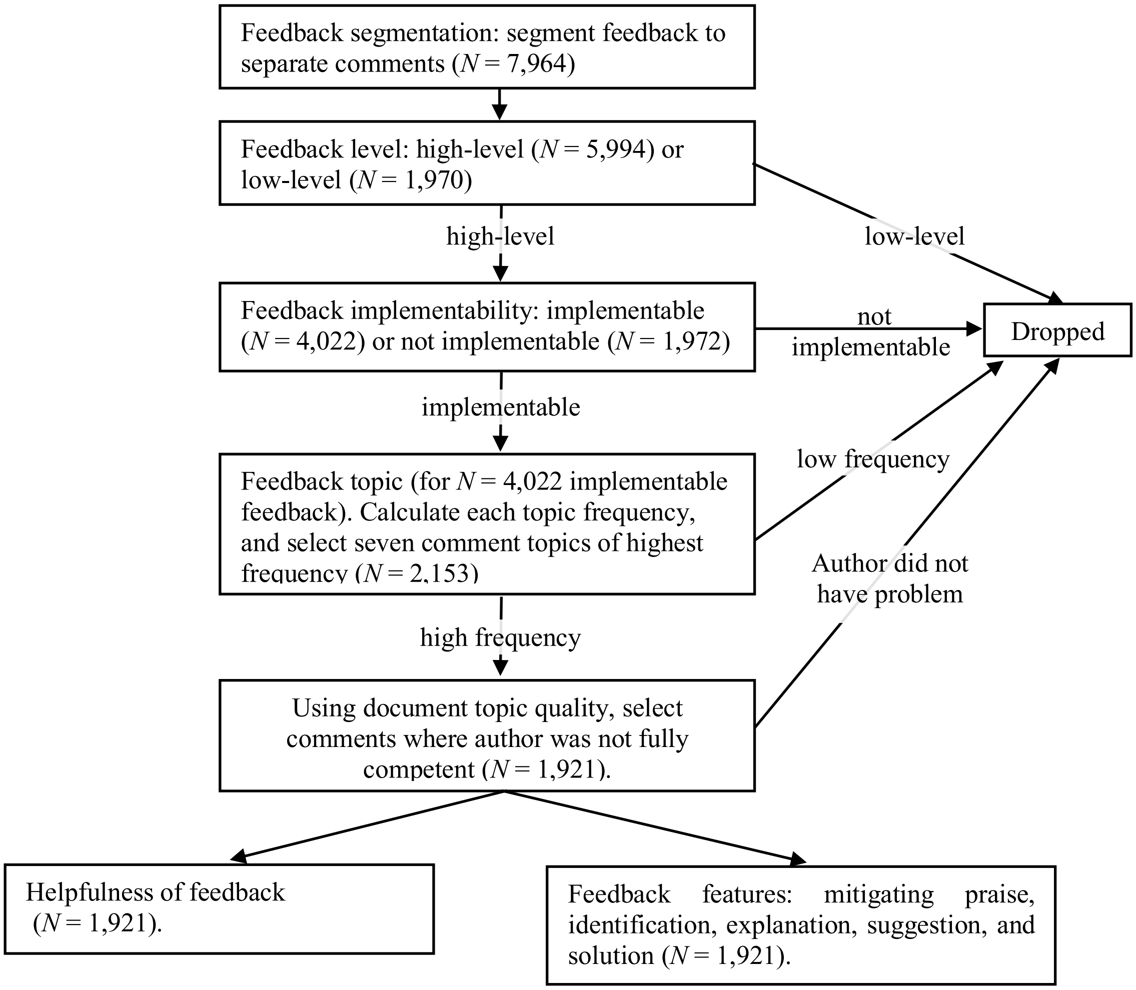

Feedback segmentation: segment feedback to separate comments (N = 7,964)
Feedback level: high-level (N = 5,994) or low-level (N = 1,970)
high-level
low-level
Feedback implementability: implementable (N = 4,022) or not implementable (N = 1,972)
not implementable
Dropped
implementable
low frequency
Feedback topic (for N = 4,022 implementable feedback). Calculate each topic frequency, and select seven comment topics of highest frequency (N = 2,153)
Author did not have problem
high frequency
Using document topic quality, select comments where author was not fully competent (N = 1,921).
Helpfulness of feedback (N = 1,921).
Feedback features: mitigating praise, identification, explanation, suggestion, and solution (N = 1,921).

Fig 3. Peer feedback coding process (see corresponding description in the section 3.3.1), along with number of comments included at each step

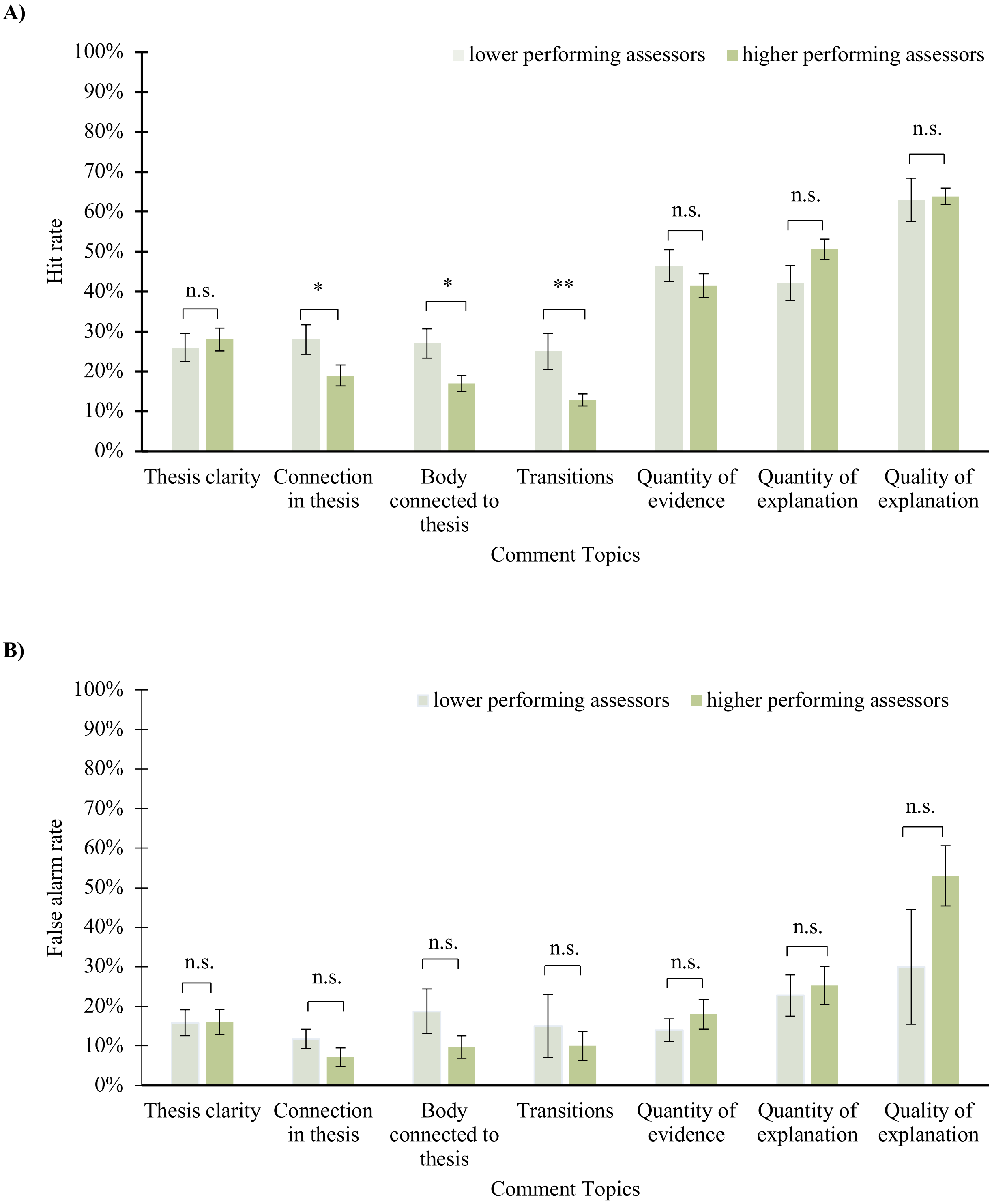


Fig 4. Mean A) hit rates and B) false alarm rates (with SE bars) of higher performing and lower performing assessors (based on topic performance) for each of the seven most common comment topics

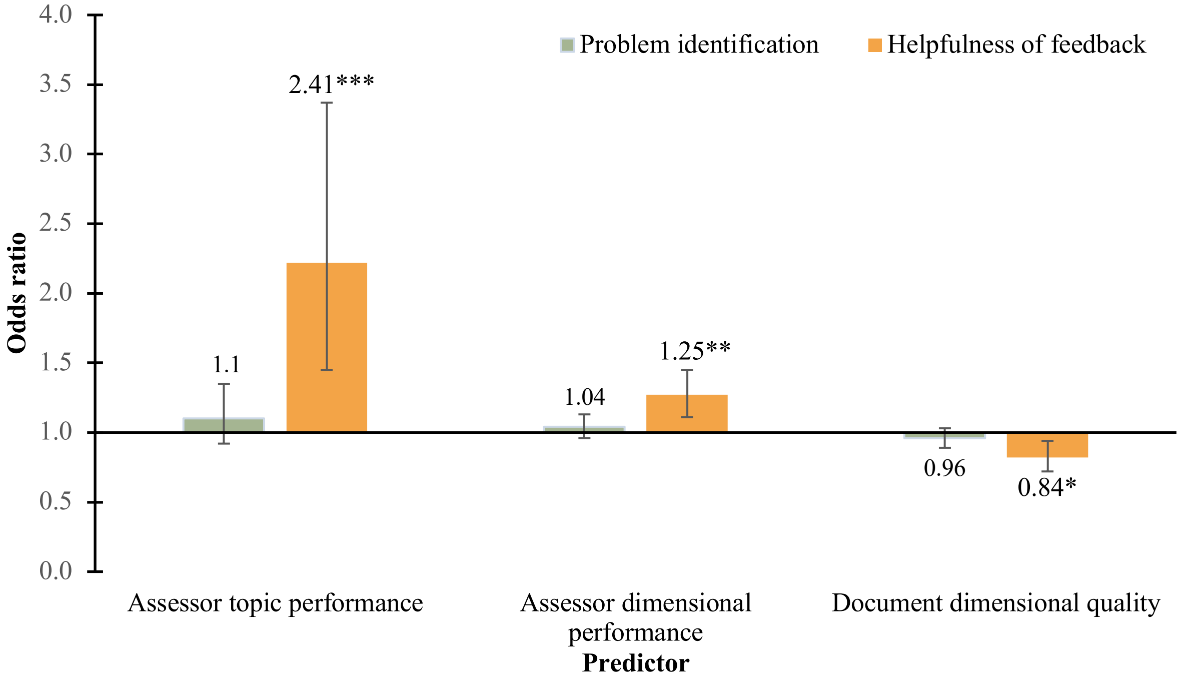


Fig 5. Effects (as odds ratios with SE bars derived from the logistic regression equations) of assessor topic performance, assessor dimensional performance, and document dimensional quality on likelihood of identifying the problem and providing helpful feedback. Note that odds ratio = 1 indicates no relationship and odds ratio < 1 indicates a negative relationship. * *p*<.05, ** *p*<.01, *** *p*<.001

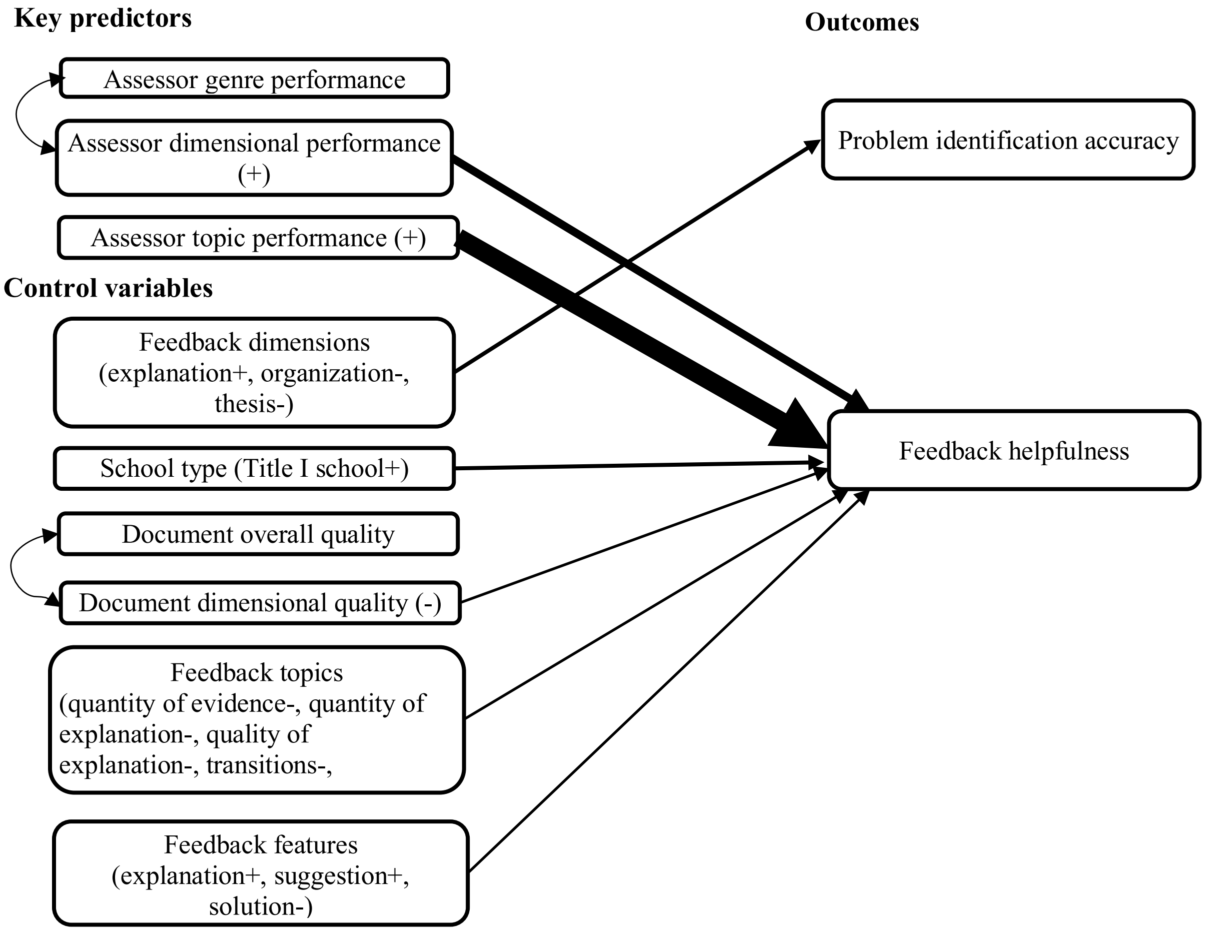

Key predictors
Outcomes
Assessor genre performance
Assessor dimensional performance (+)
Assessor topic performance (+)
Problem identification accuracy
Control variables
Feedback dimensions (explanation+, organization-, thesis-)
School type (Title I school+)
Feedback helpfulness
Document overall quality
Document dimensional quality (-)
Feedback topics (quantity of evidence-, quantity of explanation-, quality of explanation-, transitions-,
Feedback features (explanation+, suggestion+, solution-)

Fig 6. Revised model of the relationships between assessor writing performance with problem identification accuracy and usefulness of peer feedback. Line thickness corresponds with strength of relationship. Curve lines with double-headed arrows represent correlated variables. Plus/minus signs indicate positive/negative relationships. Assessor genre performance and document overall quality were not included in both models because they were strongly correlated with assessor dimensional performance and document dimensional quality respectively. Document topic quality was not included in both models because an assessor only needed to detect a problem in a comment when the assessee was not fully competent on the comment topic.

Appendix A.

Peer Review Rubrics and Comment Prompts for Each Reviewing Dimension. Note that even-value ratings could be selected, representing a value somewhere between the odd-value rating anchors.

Thesis

Comment prompt: Provide feedback on the quality of the author's thesis.

Rating prompt: 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.

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.

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

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

Argument

Comment prompt: Provide feedback on the accuracy of the author's argument.

Rating prompt: 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.

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

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

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

Comment prompt: Provide feedback on how well the author analyzed the rhetorical strategies Louv used to convey his message.

Rating prompt: 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 nosalgia, or other sophisticated strategies).

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

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

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

Comment prompt: Provide feedback on how well the author supported his or her analysis of Louv's rhetorical strategies with an adequate amount of specific and accurate references to the text.

Rating prompt: 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.

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

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

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

Explaining evidence

Comment prompt: Provide feedback on how well the author explained the textual evidence he or she provided.

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.

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

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

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

Organization

Comment prompt: Provide feedback on how well the author organized his or her essay.

Rating prompt: 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.

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.

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.

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

Comment prompt: Provide feedback on how controlled and sophisticated the author's use of language was.

Rating prompt: 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.

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

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

1 - Simplistic style and vocabulary.

Conventions

Comment prompt: Provide feedback on how controlled and sophisticated the author's conventions were.

Rating prompt: 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.

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.

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.

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 of high-level comment topics organized by writing dimensions with bolded short descriptors for the high-frequency comment topics

Dimensions	Feedback topics	Feedback examples
Thesis	1. There is no thesis in the introduction.	You should add a thesis and better topic sentences to your essay.
	2. The thesis is overly general or simply a restatement of the essay prompt. **(Thesis 1: Thesis clarity)**	Your thesis was easy to see in the passage, however it seems a little ambiguous. Louv uses devices, but how does he feel about the topic? It could help point out his view.
	3. The thesis does not make a specific or clear connection between rhetorical strategies and the author's argument. **(Thesis 2: Connection in thesis)**	The thesis describes the author's purpose. However, it doesn't describe the rhetorical strategies.
	4. There is no background information.	Thesis paragraph needs to be longer and possibly act as an intro, not just a thesis.
	5. The thesis does not align with body paragraphs.	The rhetorical devices you stated in your thesis (example, rhetorical questions, and anaphora) were a little different from your actual body paragraphs.
	6. Others (i.e., thesis problems that are not covered by the above categories)	The thesis is good at guiding the reader about what the essay will be about but it is missing complexity.
Organization	1. Organization is not clear.	The overall essay needed just a bit more organization to it, but was mostly together.
	2. Body paragraphs are unfocused.	The essay was well structured. Each passage focused on a singular topic, and mostly stayed on topic except for a few minor deviations.
	3. Body paragraphs are not clearly connected to the thesis. **(Organization 1: Body connected to thesis)**	You had good organization and all of your paragraphs followed the same organization, but you did not always connect each paragraph back to your thesis and the authors purpose for writing. You need to make sure to tie back each paragraph to your thesis.
	4. Formulaic or lack of transitions (**Organization 2: Transitions)**	You did an outstanding job on organization. Just make sure to use better transitions in each paragraph.
	5. Others (i.e., problems that are not covered by the above four categories)	You had a usual 5 paragraph essay which is good, but it seemed as if your paragraphs got shorter and shorter as you read further into the essay.
Argument	1. The argument is not addressed.	The author failed to mention Richard Louv's argument pertaining to the separation of nature and people.
	2. The argument is misunderstood.	You argued about nature vs technology rather than arguing about the separation between nature and humans, so it is a bit of a misread of the prompt.
	3. The argument is vague/incomplete.	Did not do a good job of building a strong argument to support your quotes.
Strategies	1. There is not rhetorical strategy.	There was no rhetorical strategies used at all in the essay.
	2. Analyze only obvious rhetorical strategies.	The author chose obvious rhetorical strategies (contrast, example and anaphora) to support their view on Louv's argument
	3. Rhetorical strategies are not enough.	The only legitimate rhetorical strategy that you identified was rhetorical questions, and, other than that, you simply described points in the essay. Look through the essay and try finding specific devices like ethos, imagery, anecdote, or the appeal to the reader's nostalgia.

Dimensions	Feedback topics	Feedback examples
	4. Rhetorical strategies are misunderstood.	“A warning” isn’t a rhetorical device, so you could mention that Louv gives a warning, but it shouldn’t be one of your rhetorical devices.
	5. Rhetorical strategies are vague.	The “mastery of anecdote” is a little confusing, is it related to characterization? setting? Could be a little more specific.
	6. Too many rhetorical strategies	Try to stick to a maximum of three devices if possible.
Evidence	1. Evidence is not accurate.	The quote in the first body paragraph does not accurately connect with the claim made about humans becoming consumed by materialistic items.
	2. Evidence is not complete.	Each paragraph had a direct quote (without paragraph citation) from the speech regarding the specific device.
	3. Evidence is not conveyed through direct quotes.	Good. Give a specific reference to it in the first paragraph. Need more in the second and third. Direct quotes would help.
	4. Evidence is not enough. **(Evidence 1: Quantity of evidence)**	The author barely included any evidence to support analysis. Only claims were made without evidence.
	5. Too many pieces of evidence	Though the textual evidence is relevant and well supported by the author, the paragraphs (especially body paragraph one) have too many quotations.
Explanation	1. Evidence/argument is not explained. **(Explanation 1: Quantity of explanation)**	You are not really explaining Louv’s argument. You seem to restate what Louv said. You should explain what Louv is saying. You should change your paragraphs so that they explain Louv instead of paraphrasing Louv.
	2. Explanations of evidence/argument are simplistic, or not clearly connected to the argument. **(Explanation 2: Quality of explanation)**	Thank you for explaining an anecdote with another anecdote, essentially proving nothing.

Appendix C.

Peer feedback coding scheme with definitions, examples, and interrater reliability

Category	Definition	Example
Feedback scope (*Kappa* = 0.92)		
High-level	Comments related to thesis, arguments, rhetorical strategies, organization, evidence, and explanation	The thesis contains three main points and is fairly understandable but isn't written particularly well. It doesn't follow the preferred structure of a rhetorical analysis essay.
Low-level	Comments related to use of language and writing conventions	The several misspelled words are distracting.
Helpfulness of feedback (*Kappa* = 0.70)		
Feedback with large positive effect	A comment including meaningful explanation and/or useful suggestion could improve essay quality measurably on rubrics.	You explained your references to an extent but you need to go into further depth and connect them back to your thesis and the author's purpose for writing. You only had about one sentence of explanation and it did not go into much detail (e.g., paragraph 3).
Feedback with small positive effect	A comment could improve essay quality but not measurably on rubrics.	The author did not explain the textual evidence well. The author should have expanded on his or her ideas.
No effect	A comment could not influence essay quality.	The quality of your thesis is average. It's straight and to the point. I think you need to add some descriptive adjectives.
Incorrect	A comment would harm essay quality.	Thesis is absent.
Mixed effect	A comment could improve essay quality in part, and hurt essay quality in part.	You accurately named the rhetorical devices, but just barely touched on them in each paragraph before wandering away from them.
Feedback features		
Mitigating praise (*Kappa* = 0.87)	To soften criticism in a negative comment	There is a good use of direct quotes. However, the examples are not always a true example of the rhetorical device that is being discussed.
Identification (*Kappa* = 0.78)	To identify a problem explicitly	They explained it, but there wasn't much evidence.
Explanation (*Kappa* = 0.76)	To provide explanatory information such as the nature of the problem	You went into good depth about the body paragraphs, it was well written although some parts seemed as if they did not really fit. The second paragraph was focused on pronouns but you started speaking about pathos, this point seemed to be a little out of place.
Solution (*Kappa* = 0.74)	To provide specific advice to solve the identified problem	I love how you set up your thesis as past, present, and future. The only thing I might change would be to say instead of empathy change it to an appeal to pathos.
Suggestion (*Kappa* = 0.72)	To provide general advice to solve the identified problem	Lack organization, use transition words.

Appendix D. Pearson intercorrelations among variables at the review and comment levels

Table D1. Pearson correlations among assessor topic performance, assessor dimensional performance, document dimensional quality, assessor genre performance, document overall quality, feedback dimensions, and problem identification accuracy

	Variables	1	2	3	4	5	6	7	8	9
1	Assessor topic performance	1.00								
2	Assessor dimensional performance	0.33**	1.00							
3	Document dimensional quality	0.03	0.02	1.00						
4	Assessor genre performance	0.26**	0.78**	-0.01	1.00					
5	Document overall quality	-0.03	-0.04*	0.75**	-0.02	1.00				
Dimensions										
6	Evidence	0.10**	0.10**	0.05**	0.01	-0.04*	1.00			
7	Organization	-0.07**	-0.03	0.03	0.00	0.01	-0.25**	1.00		
8	Thesis	0.16**	-0.05**	-0.12**	0.00	-0.03	-0.20**	-0.37**	1.00	
9	Explanation	-0.13**	0.01	0.04*	-0.01	0.05**	-0.26**	-0.49**	-0.38**	1.00
10	Problem identification accuracy	0.00	0.05**	0.01	0.05**	0.01	0.06**	-0.26**	-0.12**	0.32**
	M	0.26	4.68	4.39	4.83	4.60	0.12	0.32	0.23	0.33
	SD	0.44	1.03	1.00	0.76	0.73	0.32	0.47	0.42	0.47

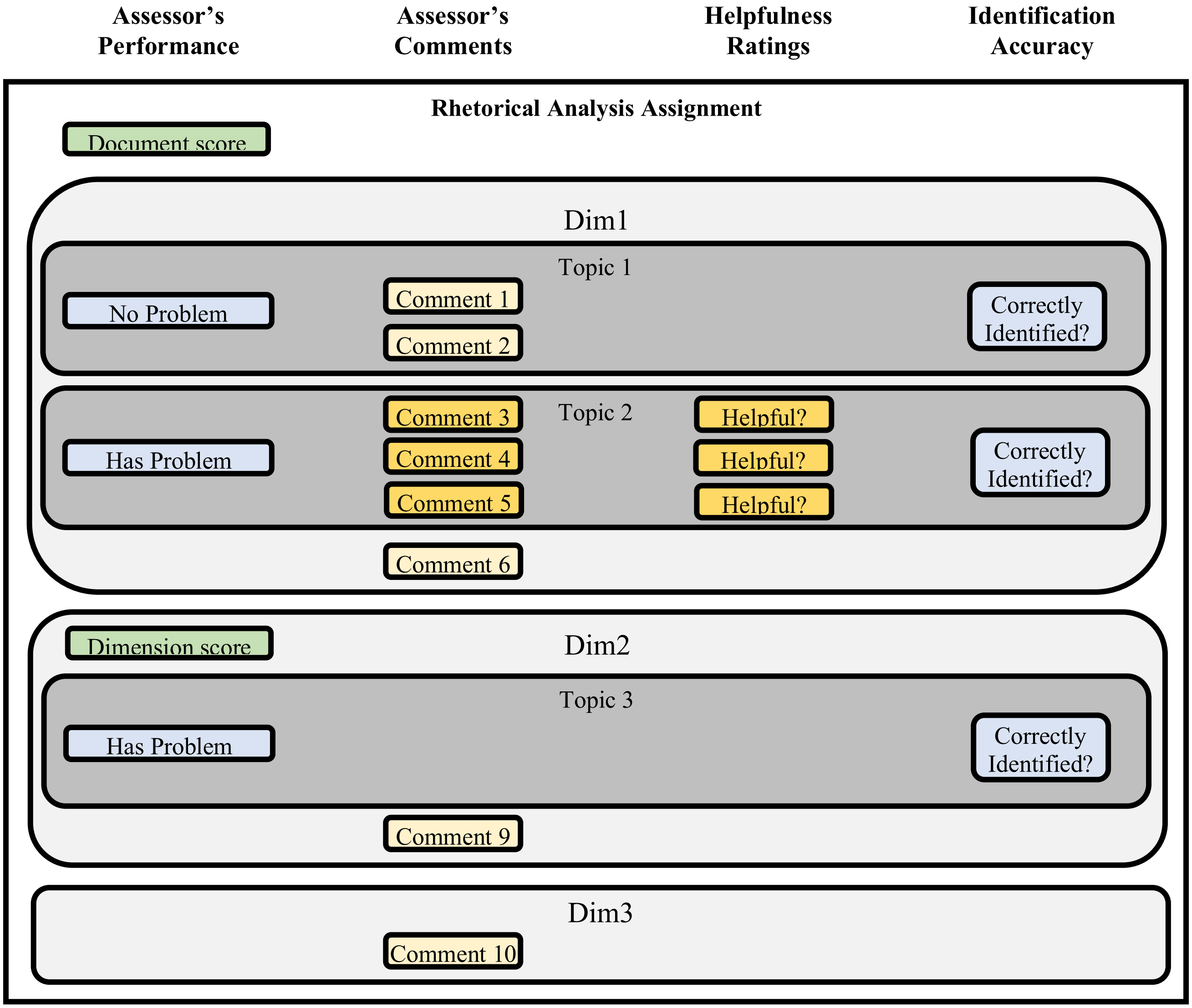
Notes. The correlation analysis was conducted with N =234 assessors. **p* < .05, ***p* < .01

Table D2. Pearson correlations among assessor topic performance, assessor dimensional performance, document dimensional quality, assessor genre performance, document overall quality, feedback features, and helpfulness of feedback

	Variables	1	2	3	4	5	6	7	8	9	10	11
1	Assessor topic performance	1.00										
2	Assessor dimensional performance	**0.28****	1.00									
3	Document dimensional quality	0.02	0.03	1.00								
4	Assessor genre performance	**0.23****	**0.80****	0.03	1.00							
5	Document overall quality	-0.02	-0.00	**0.81****	0.00	1.00						
Feedback features												
6	Mitigating praise	0.02	0.03	**0.17****	0.03	**0.19****	1.00					
7	Identification	-0.02	0.01	**-0.09****	0.03	**-0.07****	**-0.17****	1.00				
8	Explanation	-0.03	**0.06****	-0.03	**0.10****	0.00	**-0.17****	**0.49****	1.00			
9	Solution	0.03	**0.06****	0.03	**0.07****	0.02	-0.02	**-0.10****	-0.03	1.00		
10	Suggestion	-0.03	**0.09****	**0.09****	**0.06****	**0.08****	**0.09****	**-0.35****	**-0.20****	-0.01	1.00	
11	Comment length	-0.04	**0.13****	**0.13****	**0.17****	**0.15****	**0.05***	**0.19****	**0.42****	**0.29****	**0.15****	1.00
12	Helpfulness of feedback	**0.12****	**0.12****	**-0.06***	**0.13****	-0.03	-0.00	0.04	**0.09****	-0.04	**0.06***	**0.07****
	M	0.25	4.78	4.33	4.89	4.57	0.44	0.72	0.47	0.10	0.63	58.6
	SD	0.44	0.97	1.03	0.70	0.77	0.50	0.45	0.50	0.29	0.48	31.1

*Notes. The correlation analysis was conducted with N = 1,921 comments. * p < .05, ** p < .01*

Appendix E.



Notes. Problem identification accuracy and feedback helpfulness analysis process.
1. Identification accuracy was conducted at the document-topic level, not at the comment level. Therefore, every assessor contributed an accuracy measure for each problem topic (no exclusions).
2. However, feedback helpfulness was analyzed at the comment level. All comments fall within a dimension. Sometimes multiple comments from one assessor were on one topic (e.g., Comments 1-2, 3-5). Some dimensions have multiple common topics (e.g., Dimension 1) and some have no common topics (e.g., Dimension 3). It may happen that an assessor made no comments on a given topic (e.g., Topic 3). Some comments were not on a common topic (e.g., Comments 6, 9, 10). Comments were included in helpfulness analysis if they fall on a common topic (e.g., Comments 3-5), and the reviewed document has a problem on that topic.