Reviewer, Essay, and Reviewing-Process Characteristics that Predict Errors in Web-based Peer Review

Article *in* Computers & Education · January 2021

DOI: 10.1016/j.compedu.2021.104146

2 authors:

Yao Xiong
Pearson Inc
16 PUBLICATIONS **335** CITATIONS
SEE PROFILE

Christian D Schunn
University of Pittsburgh
342 PUBLICATIONS **9,747** CITATIONS
SEE PROFILE

Computers & Education 166 (2021) 104146

Contents lists available at ScienceDirect

Computers & Education

journal homepage: http://www.elsevier.com/locate/compedu

Reviewer, essay, and reviewing-process characteristics that predict errors in web-based peer review

Yao Xiong [a], Christian D. Schunn [b,*]

[a] *Pearson Assessments, USA*
[b] *University of Pittsburgh, USA*

ARTICLE INFO

Keywords:
Peer review
Review error
Essay
Reviewer
Reviewing process characteristics that predict errors in web-based peer review

ABSTRACT

Accuracy of peer review continues to be a concern for instructors in implementing computer-supported peer review in their instructional practices. A large body of literature has descriptively documented overall levels of reliability and validity of peer review and which factors across different peer review implementations impact overall reliability and validity of peer review (e.g., use of rubrics, education level, training). However, few studies have examined what factors within a peer review implementation contribute to review accuracy of individual reviews and knowledge about these factors could shape new interventions to avoid or remediate errors in particular reviews. In the current study, we tested a three-level framework (reviewer, essay, and reviewing process) for predicting the location of peer review errors. Further, we examined what factors within each level are predictive of two different types of review errors: severity and leniency. Leveraging a large dataset from an Advanced Placement English and Composite course implementing a common assignment with web-based peer review across 10 high schools, we found support for all levels in the framework and the importance of separating severity and leniency errors: review comment length predicted both severe and lenient errors but in opposite directions: longer comments are more likely to be associated with severe errors and less likely to be associated with lenient errors; review disagreement, reviewer ability and average sentence length of comments predicted severe errors; and essay quality predicted lenient errors. Implications for the development of new web-based tools for supporting peer-review are discussed.

1. Introduction

1.1. Peer review and its accuracy

Peer review is defined as an educational activity where students assess the quality of work by other students of similar status. Students of similar status are often students enrolled in the same class or in the same program, who share the context but are not expert yet in the content to be peer reviewed. Although peer review sometimes is also referred to as peer assessment, we use peer review throughout this paper to focus on the reviewing activity and its characteristics. Peer review has been widely used in both K-12 and higher education across different disciplines, and now in web-based form (Li et al., 2016; Sanchez, Atkinson, Koenka, Moshontz, & Cooper, 2017; Topping, 1998). It has been widely used for formative assessment purposes to guide student learning (Sanchez et al.,

[*] Corresponding author.
E-mail address: schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.compedu.2021.104146
Received 31 July 2019; Received in revised form 18 January 2021; Accepted 22 January 2021
Available online 26 January 2021

2017; Topping, 1998) and for summative purposes to give instructors and students summative information (Patchan, Schunn, & Clark, 2017; Suen, 2013). For example, Cho and Schunn (2007) showed how a web-based peer review system can be used by students to effectively revise papers. Peer review is popular for both the logistic reason of reducing instructors' burden and its pedagogical benefits for disciplinary content learning (Sadler & Good, 2006), for promotion of cognitive and metacognitive skills (Topping, 1998), and for enhancing social relationships and establishing trust in a learning community (van Gennip, Segers, & Tillema, 2009, 2010).

A common observation that motivates the current work is that students and instructors are reluctant to rely on peer-provided feedback or grades for formative or summative purposes due to a concern about the accuracy and usefulness of peer feedback (e.g., Kaufman & Schunn, 2011). Indeed, there are a number of reasons to be concerned with using peer reviews even in formative feedback situations: 1) students tend not to revise when they receive very high ratings (Patchan et al., 2016); 2) harsh ratings can lead to negative self-evaluations which can then produce avoidant behaviors (Elizondo-Garcia et al., 2019); and 3) several online systems hold students accountable for the rating accuracy as a pressure to take the reviewing tasks seriously (Patchan et al., 2017).

To address the concern about peer review, a large portion of peer review research, especially at the higher education level, has been focused on reliability and validity of ratings (e.g., (Chang, Tseng, Chou, & Chen, 2011; Cho, Schunn, & Wilson, 2006; Falchikov & Goldfinch, 2000; Hovardas, Tsivitanidou, & Zacharia, 2014; Li et al., 2016; Luo, Robinson, & Park, 2014; Preston & Colman, 2000; Tsivitanidou, Zacharia, & Hovardas, 2011)). Interestingly, reliability and validity of peer ratings were generally found to be at acceptable levels among those studies. A recent meta-analysis found a high average correlation between peer and instructor ratings of 0.63 (Li et al., 2016). An earlier study of 16 different higher education courses reported an inter-rater reliability among peer raters that was generally medium to high, ranging from 0.45 to 0.88 (Cho et al., 2006). However, a few other studies have found lower levels of peer reliability and validity (e.g., Chang et al., 2011; Tsivitanidou et al., 2011). The varied levels of reliability and validity may be related to how the peer review was carried out (Patchan et al., 2017; Schunn et al., 2016).

This concern about reliability and validity holds across the wide range of contexts in which peer review is applied, such as assignments/tasks/artifacts of different forms (e.g., oral presentations, written documents and reports, programming code, or design products) and with different subject disciplines. While some researchers hypothesized that peer reviews in science/engineering subject disciplines may have higher accuracy than those in social science/arts, meta-analyses showed no significant difference between science/engineering and social science/arts in terms of peer review validity measured by correlation between peer and expert ratings (Falchikov & Goldfinch, 2000; Li et al., 2016). The most prevailing factors associated with peer review accuracy were found to be on how the peer review activities were carried out, e.g., peer raters' understanding about the rating rubrics, and whether peer reviewers and reviewees were matched at random (Li et al., 2016).

1.2. Peer review accuracy at macro and micro levels

Most importantly here, even a high overall accuracy of peer review results (e.g., a correlation of 0.7 between peer and expert ratings) can mean that a non-trivial number of documents have received inaccurate grades, and indicators of which documents are likely to have been incorrectly graded are important to develop. With such information, new systems could be developed that automatically discount certain ratings, assign documents to additional reviewers, or flag reviews requiring further evaluation by instructors or teaching assistants.

To date, peer review reliability and validity issues have been well documented within what we term the *macro-level* lens. Peer review accuracy at the *macro-level* lens is defined as measurement of peer review accuracy at the level of assignment or higher, e.g., peer review accuracy in a course measured by correlation between peer ratings and instructor ratings. In those cases, one statistical number (e.g., Pearson's r) could represent accuracy in a course involving many peer reviewers/reviewees participating in a peer review activity. Studies addressing peer review accuracy at the macro level usually tackles two types of research questions: 1) what the overall reliability/validity is for the peer review implemented under certain contexts (e.g., descriptive studies: Chang et al., 2011); 2) what affects the overall reliability/validity across different peer assessment implementations (e.g., course content, course level, face-to-face vs. online reviewing, assignment type, rater training, rubric explicitly: Falchikov & Goldfinch, 2000). The second type of research are usually meta-analyses that synthesizing many studies under the first type of research (e.g., Sanchez et al., 2017). Despite wide variation in methods across those studies, the common thing is that peer review accuracy is considered as a property of the whole peer review activity, rather than analyzing variation in accuracy at the level of individual reviewer, reviewee, or review.

While research in a macro-level lens can help system designers and instructors arrange for higher overall validity of scores in the courses implementing peer review, some individual documents within a peer review task will inevitably still be incorrectly scored and instructors will seek support in addressing those mis-scored documents. It is, therefore, important to investigate what factors are associated with the accuracy of reviews of a specific document provided by a specific reviewer. This individual review level of accuracy, which we term the peer review accuracy at the *micro-level*, has been rarely studied. Peer review accuracy at the *micro-level* is defined as measurement of peer review accuracy at the individual review (or reviewer or document) level. A study involving peer review accuracy measured using a micro-level lens focuses predominantly on features that vary at the micro-level (e.g., the characteristics of the document being evaluated, the reviewer, or the review itself). A study at the macro-level might consider averages in those same features (e.g., the general characteristics of the pool of documents or reviewers), but macro-level will also consider features that can only vary at the larger grain size (e.g., features and general parameters of the assignment, level or discipline of the course, overall class/school climate).

Some recent studies that have reported peer review results with a micro-level lens mainly focused on the cognitive aspects of peer review process, e.g., examining what characteristics in peer feedback (e.g., directive/non-directive, global/local, and presence of solution or explanation in the feedback) were associated with student acceptance of the feedback, implementation of the feedback in their revisions, and quality of their revisions (e.g., (Cho & MacArthur, 2010; Gao, Schunn, & Yu, 2019; Patchan & Schunn, 2016; Patchan, Schunn, & Correnti, 2016; Saeed & Ghazali, 2017; Wu, Petit, & Chen, 2020)), or focused on the qualitative differences between peer and expert comments (e.g., Wu, Petit, & Chen, 2015). Many interesting findings were revealed. Specifically, several characteristics of peer feedback were associated with improved higher-level revision, e.g., non-directive comments, global-level comments, presence of solution, explanation and hedges in the comments, and mitigating praise in the comments (Cho & MacArthur, 2010; Saeed & Ghazali, 2017; Wu et al., 2020).

However, prior studies including micro-level measures mainly focused on cognitive or qualitative aspects of peer review comments, instead of investigating the quantitative discrepancy between peer and expert reviews or peer review errors. We only found one recent study that tackled peer review accuracy at the micro-level lens, examining occurrence of lenient and severe errors (a micro-level measure) (Liu et al., 2019). Interestingly, this study investigated how macro-level contextual variable, peer review requirement: compulsory vs. voluntary, affected micro-level lenient and severe errors in peer ratings.

By further investigating accuracy attached to individual peer reviews (particularly the process by which peers interact with documents during peer review), can we uncover the factors that are associated with those errors and eventually build interventions to efficiently address errors that do occur and further improve accuracy of all individual reviews. After all, the goal of formative assessment in education is to provide fair assessment to each individual student.

2. The framework for micro-level sources of review errors

2.1. Review accuracy and errors at the micro-level

The aim of this study is to examine peer review errors, the opposite of peer review accuracy. In peer review contexts, accuracy normally refers to the agreement between peer reviews and expert reviews (AERA, APA, & NCME, 2014; Cho et al., 2006; Li et al., 2016). Therefore, review error, defined here as the discrepancy between peer reviews and expert reviews, is a threat to validity in which expert reviews are treated as the "gold standard" (AERA et al., 2014). By contrast, reliability-related errors are discrepancies within peer raters, such as errors related to intra-rater reliability (e.g., inconsistency within one peer reviewer from assignment to assignment), and errors related to inter-rater reliability (e.g., disagreement among different peer reviewers on the same piece of writing) (Suen, 2014).

Sources of error have been examined in many contexts. The current study was carried out with writing classes using web-based peer review in a diverse range of secondary schools; peer review is often included in such contexts, but peer review research of these contexts has been rare (Li et al., 2016; Sanchez, Atkinson, Koenka, Moshontz, & Cooper, 2017; Schunn, Godley, & DeMartino, 2016), and the overall accuracy has sometimes found to be acceptable (Schunn et al., 2016), but at other times not acceptable (Hovardas et al., 2014). Given the scarcity of peer review research conducted in high schools overall, we did not limit our literature review below to be within this specific context, but attempted to build our framework based upon a broad range of peer review research to develop a comprehensive account.

Theoretically, review errors can be divided into those caused by the *macro-level* context, contextual conditions that apply to all the peer reviewers working on a given assignment in a given course (e.g., ambiguities of the writing task, average skill level of students in the course, ambiguities in the reviewing instruction, scaffolds for reviewers in the system or by the teacher, training for the reviewers) and those caused by the *micro-level* context, features that pertain to a specific reviewer, document, or review within a peer review implementation (e.g., relative ability of each reviewer, and characteristics of the reviewed document within a particular macro-level context). Here we discuss only the micro-level, although the framework can serve as a foundation for thinking about the effects of the

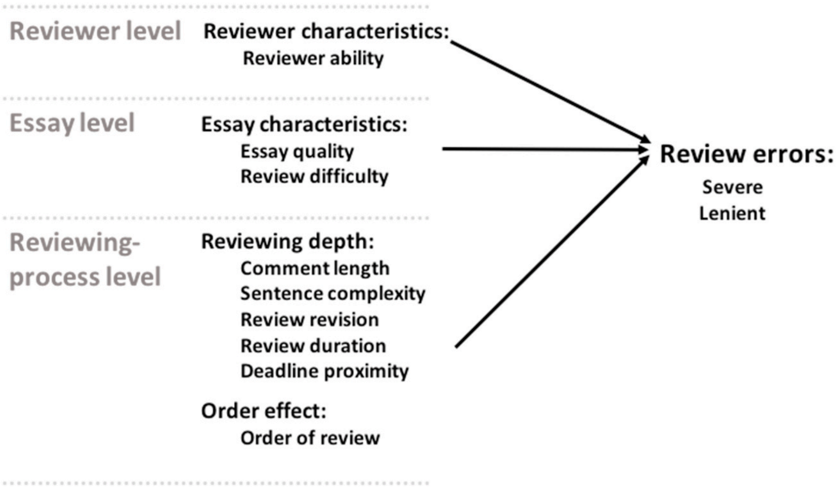

Fig. 1. The framework for reviewer, essay, comment, and reviewing-process characteristics that are predicted to influence review errors.

macro-level.

Since peer review should be theoretically conceptualized as a dyadic process between an individual and an object, errors in assessments by a particular reviewer in assessing a particular document can come from characteristics of the reviewer (e.g., lack of skill for the task, lack of motivation to review, biased expectations), characteristics of the document (e.g., obvious or subtle problems), or characteristics of the reviewing process (e.g., conducted under ideal or stressed/tired/otherwise sub-optimal circumstances). While the characteristics of reviewees (or authors) may also play a role in the peer review process (e.g., in the revision process after receiving peer review), we did not specifically include this level in the framework due to two reasons: 1) the current study focused on the peer review process when reviewers directly interact with the reviewees' artifacts (i.e., essays), but not with the reviewees *per se*; and 2) the reviewee characteristics are essentially at the same measurement level as the document (i.e., essay) level already included in the framework. The reviewing-process characteristics are related to both the reviewer and the document, when the interaction happens of a specific reviewer reviewing a specific document. At the same time, the reviewing-process characteristics are also specific in their own because the reviewing process of a specific reviewer on a specific document is not solely dependent on the characteristics of the reviewer and the document, but more on the interaction of the two.

Fig. 1 presents a summary of the three-level theoretical framework regarding errors in web-based peer review. Other levels can exist within a class, such as the assignment level when multiple assignments are given sequentially (i.e., peer review errors can be measured within assignments 1, 2, or 3), or the reviewing dimension level when the reviewing rubrics for an assignment requires different ratings for different reviewing dimensions. However, we did not include those specific levels in the current framework mainly because we attempted to build a more generalizable three-level framework that includes commonly measurable factors in the web-based peer review context.

The framework includes the distinction between rating errors as being either too severe or too lenient, where leniency refers to the error of assigning higher ratings than the actual quality of the assignment and severity referring to the opposite error direction. A number of studies of peer review reliability have noted that peers tend to give ratings that are too high (e.g., Matsuno, 2009), but errors do occur in both directions. Most importantly, theoretically-different factors will often underlie severe vs. lenient errors (e.g., related to errors of commission vs. errors of omission), as described in the next section. Therefore, we examine the role of each source to each error type.

Determining which level of the three-level possible factors contributes most to the peer review errors is important for both theoretical and practical reasons. Theoretically, the three-level framework provides a general framework to study peer review using the micro-level lens, and identifies potential sources of peer review errors, which then can be fed into larger models of learning from peer review. Pragmatically, it reveals primary aspects that need support and indicators that new tools can leverage for addressing these problems. For example, based on reviewer-level characteristics, it could be possible to automatically assigning pools of reviewers to each document that include reviewers more likely to be accurate. Alternatively, using document or review-level characteristics, it could be possible to build automatic assignment of extra reviewers to documents with likely rating errors.

2.2. Reviewer characteristics connected to review error

Reviewer ability. The characteristics of the peer reviewers providing the review ratings has long been named as a likely source of review errors (Popham, 1989): low-performing students are often perceived to be likely sources of inaccurate reviews. In particular, reviewer ability, defined as the peer reviewer's level of skill and content knowledge related to the essay being evaluated, might influence detection of problems in writing or diagnosis of their relative importance. Reviewer ability here refers to the "relative reviewer ability" (i.e., actual ability for each individual reviewer on the specific review task in comparison to others in the class), in contrast to average reviewer ability in the class which is a macro-level measure. For simplicity, we use "reviewer ability" throughout the document referring to the micro-level individual reviewer ability.

Reviewer ability is the most commonly-studied reviewer characteristic in peer review contexts, partly because reviewer ability appears to be the major discrepancy between a novice peer reviewer and an expert instructor reviewer and also because reviewer ability is commonly measurable within a peer review system (e.g., (de Alfaro & Shavlovsky, 2016; Huisman, Admiraal, Pilli, van de Ven, & Saab, 2018; Matsuno, 2009; Patchan, Hawk, Stevens, & Schunn, 2013; Patchan & Schunn, 2016; Piech et al., 2013)). Interestingly, several of these studies did not observe a strong effect of reviewer ability on review quality in terms of review accuracy (Matsuno, 2009) or review usefulness (Patchan et al., 2013; Patchan & Schunn, 2016). Instead, to the extent there were differences, students of varying abilities appeared to focus on different aspects of writing, with no overall increase in accuracy. In the context of a massive open online course (MOOC) with its corresponding extreme variation in student ability, only the bottom 20% of peer reviewers had weaker review accuracy (de Alfaro & Shavlovsky, 2016). We include reviewer ability in the current study to explore its effect in the peer review in a high school course, given the previous literature has not examined many different contexts so that replication is important. Further, since peer assessment research at the secondary level is relatively rare, instructor and student concerns about the accuracy of lower ability peers' reviews are commonly voiced (Langer & Applebee, 1987) and need to be directly studied.

It is important to note however that the question of reviewer ability effect on review accuracy is still open. First, accuracy was not consistently defined across studies. For example, in both the Piech et al. (2013) and Matsuno (2009) studies, the authors relied on peer ratings or a combination of peer, self-, and teacher ratings to estimate the ground truth, which potentially introduced a number of biases. While de Alfaro and Shavlovsky (2016) defined accuracy as the discrepancy between peer and instructor ratings, they did not have multiple instructor ratings, and so the estimate was likely noisy. In addition, the MOOC contexts of the Piech et al. (2013) and de Alfaro and Shavlovsky (2016) studies likely had atypically large ability variation. Therefore, more research needs to be done on this topic in typical courses with multiple experts providing ratings to provide a strong ground truth.

2.3. Essay characteristics connected to review error

Essay quality. Essay characteristics have also been studied in the peer review literature, especially overall essay quality. Matsuno (2009) reported that peer reviewers tended to rate high-quality essays lower and low-quality essays higher, regardless of the reviewers' own abilities. This coincides with a central tendency effect in which the middle categories of a rating scale tend to be overused (Myford & Wolfe, 2003). In a peer review setting, it is potentially more difficult for reviewers to recognize those extremely good or poor essays given they are only provided a handful of essays to grade and thus these extreme cases may influence their expectations for the norm. However, in the context of greater ability variation within a MOOC, the opposite effect of errors, exaggeration, was observed (Piech et al., 2013), perhaps due to the variability in student ability. Indeed, the most severe reviews came from high-ability reviewers grading low-quality assignments (Piech et al., 2013).

Review difficulty. Another essay characteristic that may lead to review errors is review difficulty: some essays may be more difficult to be graded than others (Gao et al., 2019), perhaps due to uneven quality across the essay, novel writing moves, or writing issues not explicitly covered in the assignment description or reviewing rubrics. Because of the various underlying sources, review difficulty may be measured indirectly. For example, essays that involve larger disagreement across reviewers may have some specific characteristics (e.g., including novel writing moves) that led to the review difficulty. Review disagreement was previously found to predict review accuracy (Nguyen & Litman, 2013b), with more disagreement linked to lower accuracy.

2.4. Reviewing-process characteristics connected to review error

Reviewing depth: Text comment. As part of the process of generating ratings, most peer review implementations include a step of generating free text comments; such inclusion generally improves the accuracy of the ratings (Li et al., 2016). The accuracy of the text comment is likely a strong predictor of the accuracy of the numeric ratings (Rico-Juan, Gallego, & Calvo-Zaragoza, 2019) because they are both connected to the central objective of the peer reviewing process: detecting and diagnosing problems. However, accuracy of the comment has many components, such as accuracy of detection, accuracy of diagnosis, and accuracy of advice, only some of which are related to accuracy of the rating. For example, logically speaking, accuracy of detection and accuracy of diagnosis as the foundation of the ratings and thus closely related to its accuracy, whereas accuracy of the advice involves additional knowledge/information that is at best indirectly correlated with the accuracy of the rating because the ratings are generally about the frequency and severity of problems in the document. Further, there can often be many comments associated with one rating, and the accuracy of the comments as a collection will be difficult to assess.

A more basic process-oriented dimension of review comment is the overall depth of the comment, which has been previously connected to rating reliability (Patchan et al., 2017). Based on analogy research on writing quality (Beers & Nagy, 2009), longer comments with more complex sentences likely reflect more extensive reasoning, which might be associated with more accurate evaluation. Longer comments tend to include a number of useful comment details, which are generally associated with greater document improvements (Patchan et al., 2016). However, very long comments may reflect unrealistically high expectations on the part of a reviewer, and thus could predict ratings that are too severe. Alternatively, long comments could also involve reviewers articulating confusion, which may be associated with more errors, either lenient or severe. Therefore, the relation between characteristics of the textual comment and errors in the numeric ratings is open and exploratory.

Reviewing depth: Review revision. Another indicator of depth of the reviewing process is the amount of revision of the review comments. Since revision during writing generally tends to produce better documents (Cho & MacArthur, 2010), it is possible that revision to reviews will produce better reviews. On the other hand, more revisions on reviews can reflect confusion, which would be associated with more errors. Therefore, the number of revisions to a specific review prior to final review submission is included in the framework as well at the reviewing-process level. However, similar to the characteristics of the textual comment, this relation is also open and exploratory given the possibility of effects in both directions.

Reviewing depth: Review duration. Besides comment characteristics and revisions, other reviewing-process characteristics are potentially related to review errors. For example, timing information has often been used to provide critical insights into decision making quality in psychological research (Smith & Ratcliff, 2004). In the context of MOOCs, one study found that a slightly lower than average mean time spent in reviewing was associated with *higher* review reliability (Piech et al., 2013), and a second MOOC study found that the quickly-performed reviews (i.e., reviews completed in the first 10% percentile) actually tended to be slightly *more* accurate (i.e., 10% fewer errors) than the later reviews (de Alfaro & Shavlovsky, 2016). However, students taking too little time to read and find problems in a document (and thereby produce ratings that were too lenient) could be indicated by very short reviewing times.

Reviewing depth: Deadline proximity. Another process factor in peer review related to depth involves the timing of reviews relative to the reviewing deadline. Interestingly, reviews submitted close to the deadline within a MOOC tended to be slightly less accurate than earlier ones (de Alfaro & Shavlovsky, 2016). It may be that last-minute reviews are likely to be completed in a rush and therefore involve errors.

Order of review. Timing also involves another general factor in decision-making quality: order effects, which involve both positive (practice effect) and negative (fatigue) elements. In particular, later reviews are completed by students who have had more practice with the reviewing process and rating rubrics but are also potentially fatigued. Practice or fatigue effects, measured by order in which reviews were completed, were found to be very small in a MOOC (de Alfaro & Shavlovsky, 2016). However, this effect might differ in the context of non-volunteers or high school classes. A light reviewing load (e.g., 4 or 5 reviews) for volunteer or adult learners might not seem light to adolescents or non-volunteer learners.

2.5. Research purposes

The main research purpose of this study is to examine web-based peer review errors and possible factors commonly measurable within web-based peer review systems that can be used to understand the causes of and predict/mitigate review errors. We included factors across three levels that are potentially associated with review errors, namely reviewer characteristics, essay characteristics and reviewing-process characteristics. Examined within a high-school peer review activity in an AP English and Composition course, the specific research questions are as follows:

1. What characteristics from reviewer, essay, reviewing-process levels underlie peer review errors?
2. Are the same characteristics important for both leniency vs. severity errors?
3. Do factors within all three levels of peer review (reviewer, essay, and reviewing process) predict errors?

3. Methods

3.1. Participants

Participants included in the study were 818 students from ten different secondary schools. They were taking the same Advanced Placement course in writing (AP Language and Composition) taught by ten different teachers, each teaching between two and five different sections of this course at their school. This course has the largest annual enrollment among AP courses, with over 500,000 students enrolled in the course each year. Among the ten schools, four were Title I schools (high rates of low-income families), which accounted for 33% of the participants. Title 1 schools tend to have less-experienced teachers, and students with lower family income, lower parental education levels, and less experience with AP courses (Institute of Education Sciences, 2007). The schools were broadly distributed across the US in 9 different states. All but one of the 10 teachers were female, and all but one had taught this class for at least two prior years. Half of the teachers had taught this class using the web-based peer review system in the year before.

Student age ranged from 16 to 19 years old with a mean age of 17.3. Sixty one percent of students were female. The race/ethnicities of the students were: 69% Caucasian, 16% Asian, 8% African American, and 7% Hispanic/Latinx.

Among the 818 students, 293 student essays were sampled (approximately 30 essays per instructor) to receive expert grades. In addition, each student essay was peer reviewed on average by four peer students. Those expert-graded 293 essays, receiving 1,138 peer reviews from 620 unique student reviewers, were the sample analyzed in this study. The 293 authors of these essays, the 620 reviewers, and the full set of 818 students had very similar demographic composition.

3.2. Peer review procedures

The AP course was a face-to-face course conducted in different high schools while the peer review was conducted mainly online. A web-based writing and peer review system, SWoRD, was used by students in the current study to submit and review essays within their classes (Cho & Schunn, 2007; Schunn, 2016). As part of the class, the students were given a one-page persuasive writing passage. They were instructed to read the passage carefully and then, in a well-developed essay, analyze the rhetorical strategies the author of that passage used to develop the main argument. Students were told that they needed to support their analysis with specific references to the text. This kind of writing task is a core component of the AP course, and the particular task was selected from a prior year's high-stakes end-of-course assessment. Students were required to support their analysis with specific references to the passage and explanations.

The peer review was conducted in several shared steps across the different classes. First, instructors were provided shared assignments, shared peer review rubrics, and training on the use of the process and system. Second, students submitted their essays using the web-based peer review system before a specified deadline using assigned pseudonyms. Third, students were randomly assigned to evaluate four essays of their peers from the same class, using set of rubrics shared across all classes. Fourth, an in-class discussion at the beginning of the reviewing period was used to provide training on the peer review task and the peer review system. During the training session, the teacher shared two sample essays with all the students. Students read the first sample essay, and then were shown example reviews for it that were generally unhelpful vs. generally helpful (e.g., specific and constructive) and discussed as a class what made reviews helpful. Then students read the second sample essay and completed a review with a partner in class using the assigned reviewing rubrics. The class as a whole discussed the feedback and ratings that were generated. At this point, the rating scales used in reviewing were discussed and students received calibration feedback through the in-class discussion. Fifth, students completed their four peer reviews online and anonymously. These reviews were then made available to essay authors to guide their essay revision process. To increase the quality of the reviews (Patchan et al., 2017), students were held accountable by receiving a grade for the accuracy of their ratings (based on being consistent with other students) and for the helpfulness of their comments (based on helpfulness ratings made by the essay authors). By using random assignment of reviewers to essays, anonymity of reviews, and reviewer accountability mechanisms, students were motivated to participate seriously in this peer review activity and to provide fair reviews to their peers.

The current study focused on the online review process of the first draft of the four essays they were assigned. The peer-review activities were coordinated to be the same across the ten different classes for the first draft to better zoom in on within review factors (i.e., not confound factors studied here with varying nature of the review assignment across schools). Peer review on the second draft was not mandatory within the study and was not always implemented across instructors. Therefore, we focus the study on peer review for the first draft.

3.3. Peer review rubrics

In the peer review system, student reviewers were provided a review form to evaluate essays by giving both numerical ratings (on 7-point scales) and text comments on each of five different evaluation dimensions. These dimensions were created by the experimenters, starting with the holistic rubric used by expert ratings of the high-stakes end-of-course exams, and then dividing that holistic rubric into separate components and changing to more student-friendly language for the dimension descriptions and rating anchor points based on feedback from teachers and students (Schunn et al., 2016). In the current study, five different dimensions were examined, including 1) rhetorical strategies, 2) evidence for claims, 3) explaining evidence, 4) organization, and 5) control of language and conventions. For our analysis purposes, the first three dimensions were related to argument while the last two dimensions were about use of language. Those five evaluation dimensions (with shared scoring rubrics) were the same for all ten instructors. In addition, those five dimensional scores have medium-to-high correlations, which range from 0.61 to 0.74 for peer ratings and 0.42 to 0.69 for expert ratings. Expert graders are explained in detail in the next section.

All of the rating scales are provided in Appendix A. In the scoring rubrics, scores were corresponded to expected quality of each dimension in details. For example, in order to award 7 points on the rhetorical strategies dimension, a reviewer needs to find evidence of a student being able to analyze multiple and subtle rhetorical strategies that the author used in the prompt passage. Note the inclusion of concrete anchors (e.g., Level 7 for the *Evidence for Claims* dimension had the anchor: *Every claim has accurate evidence for all important aspects of the claim. Most evidence is conveyed through direct quotes*.). These anchors support absolute rating accuracy. The anchors were provided for the odd-numbered rating levels; even levels had no separate anchor content, a design strategy which allows for intermediate ratings without overwhelming students with 7-levels of detailed anchor text.

3.4. Ground truth: average expert ratings

In order to have a measure for ground truth, a group of expert graders are usually used (e.g., Li et al., 2016; Suen, 2014). To serve as expert graders, seven advanced graduate students studying Rhetoric and Communication at a highly ranked research university were recruited. All had taught the first-year writing course for multiple years at their university; this particular university course was the one that the AP Language and Composition course was meant to replace. Thus, these graders were experts both by research focus on rhetoric and communication and by experience in relevant instruction and grading. Further, they received multi-hour, in-person training on the rubrics. During the training, they practiced grading on a sample set of essays and received feedback on their grading performance.

Among the 818 student essays, 293 (30 per teacher unless there were fewer than 30 students for a teacher) were randomly selected to be expert-graded, which was a stratified random sampling approach with each teacher being a stratum. This number was selected as a balance between obtaining a sufficiently representative sample and the amount of work needed for expert rater training and grading.

Every essay was randomly assigned to two experts initially. All essays receiving more than a one-point difference in ratings were assigned to a third grader. As a result, 53 essays received ratings from a third grader. For those 53 essays, we retained the two ratings that had the least distance by dropping the ones that were most different, which was similar to the "Olympic average" approach.

Further, coding drift (i.e., systematic patterns by an expert grader in higher or lower ratings for a dimension across essays) was regularly assessed throughout the coding process and discussed whenever it was detected.

The mean ratings across two expert graders were considered as "ground truth" for the current study. Reliabilities of these mean ratings across the seven expert graders were calculated using an aggregate consistency-type intra-class correlation (ICC) (Cho et al., 2006), resulting in estimated ICCs of 0.83 for composite ratings (i.e., average ratings of all dimensions), and ICCs ranging from 0.65 to 0.79 for the five dimensional ratings. The higher ICC for composite ratings is consistent with the reliability theory that combining several correlated measures is usually more reliable than using a single measure. These reliability levels are higher than those typically found for instructor evaluations of writing (Cho et al., 2006), the most common benchmark for peer review studies.

The 293 essays received 1,138 peer reviews from 620 unique student reviewers. On average, one essay received 4 peer reviews. The 293 randomly selected authors/essays to be studied were representative of the 818 students in terms of essay quality. The average essay quality measured by peer ratings was 5.33 out of a 7-point scale for all 818 essays while that was 5.38 for the 293 essays. In addition, the distributions of expert ratings were not skewed (absolute skewness ranging from 0.1 to 0.3 for the five dimensional ratings).

3.5. Measures

Review errors. Review errors were calculated using the difference between student and mean expert ratings on a given essay. First, the difference scores were centered around the dimension means (i.e., separately for each dimension) in order to remove any general tendency for student reviewers to give higher ratings than experts on particular dimensions or overall (i.e., a population-wide bias is a different phenomenon from individual rating errors). Another practical reason to center the difference scores was to correct the skewed distribution of the scores (i.e., much more positive scores than negative scores), given the distributions of "ground truth" (i.e., expert ratings) were not skewed. Fig. 2 (left) shows the distribution of continuous review errors (leniency) at the composite level. We used categorical review errors (i.e., lenient error or severe error) as our dependent measure instead of continuous errors for two reasons: 1) an effect on the continuous error measure could be canceled out if a predictor is predicting both leniency and severity; and 2) a continuous measure of errors may show consistent but very small errors of little pragmatic importance. Therefore, review errors were further categorized using −1 and 1 as cutoffs: errors below −1 were categorized as *Severe*; errors above 1 were categorized as *Lenient*; and everything between (including) −1 and 1 were categorized into the *Accurate* category. We used more than 1 as the cutoff criterion for several reasons: 1) 1 point was the smallest mistake possible on the 1-to-7 rating scales; and 2) more than 1 point in composite rating was a large change in quality within the observed distributions of ratings. Fig. 2 (right) is the bar chart for the three categories. This categorical review error was the primary dependent variable in the study.

Predictors of review errors. The predictor measures were gathered from three different data sources and calculated at three different levels. In terms of sources, peer ratings were used to calculate characteristics of essays and reviewers, server log files were used to calculate review timing, and peer comments were used to determine comment characteristics. In terms of measurement levels, measures represent characteristics of reviewers, essays, or reviewing process.

Table 1 presents detailed information on all measures. Essay quality and review disagreement are essay-level measurements, and reviewer ability was a characteristic of the reviewer. The remaining measures were at the reviewing-process level. Deadline proximity, review duration, order of review, and review revision were extracted from the log files. Comment length and average sentence length (measuring sentence complexity) were calculated from the text comments. We originally calculated comment length in three ways: number of characters, number of words, and number of sentences. However, the correlations among those measures were approximately 0.9. We thus only used word counts as the comment length measure. We binned the four review timing variables as shown in Table 1 (i.e., Deadline Proximity, Review Duration, Order of Review and Review Revision) for the purpose of more meaningful and sanitized measurement. For example, when we used the logged time (i.e., time duration between downloading a paper and submitting a review) to measure review duration, it is a noisy measure of review duration. However, we are fairly confident that reviews with less than 10 min duration time were likely speeded due to the complexity of the peer review task after we examined how long our expert graders needed to generate ratings for one document. Using 10 min as a cutoff reduced the false positive rate (i.e., flagged as speeded when it is not) of the measure, although re-analyzing the data using 20 min as the cutoff produced similar findings. Using 10 min as the cutoff, approximately 4% of the reviews that were marked as speeded (see Table 2). In addition, Deadline Proximity was categorized into three levels: late submission, last-minute submission (with last 30 min), and early submission. Order of Review was consistent with the raw measure of the review order: 1st, 2nd, 3rd, and ≥ 4th. Review Revision was also categorized into two levels: no revision vs. revision. No revision means the reviewer submitted the review without further revision/modification to ratings or comments, while revision means that reviewer revised/resubmitted the review after they first submitted the review. Multiple revisions potentially mean the reviewer might have invested more thinking or reflection in the review they submitted; given the distribution was skewed (i.e., only a few with multiple revisions), we used two categories to distinguish potentially different amounts of cognitive processing. All the measures were calculated at each dimension level as well as at a composite level, except the four review timing variables, which were the same across different dimensions.

Reviewer ability was conceptually defined in terms of writing ability for the writing skills directly assessed in the assignment, in contrast to skills unique to reviewing (e.g., giving of well-worded or persuasively worded advice). It was operationally measured by their own document quality, as is commonly implemented in research on peer review (e.g., Patchan & Schunn, 2015; (Patchan et al.,

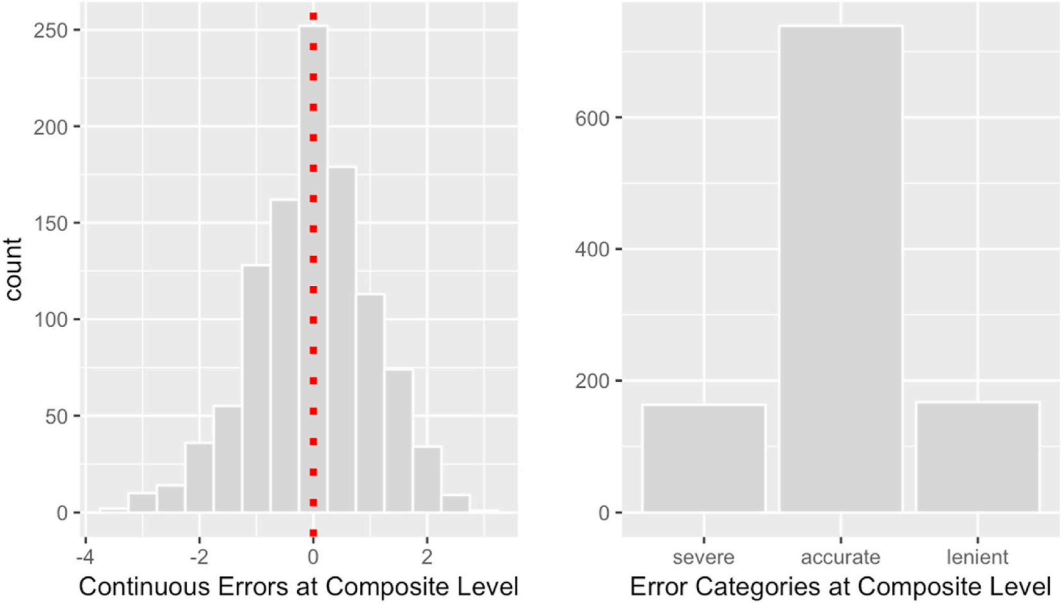

Fig. 2. Distribution of review errors.

Table 1
Summary of measures.

Measure	Measurement Level	Details
Review error	Review	Discrepancy between peer and mean expert ratings; three-level categorical variable: severe, lenient, and accurate
Essay quality	Essay	Quality of essay; average across median peer ratings and mean expert ratings
Reviewer ability	Reviewer	Ability of reviewer evaluated by student reviewers; median of peer ratings on a reviewer's own essay
Review disagreement	Essay	Disagreement of reviewers on one essay; standard deviation of peer ratings on one essay
Deadline proximity	Review	Minutes to deadline for the first submission; categorized to three levels: late submission, last-minute submission (with last 30 min), and early submission
Review duration	Review	Minutes between requesting a paper for review and submitting the review; categorized to two levels: speeded ($\leq$ 10 min) and normal (>10 min)
Order of review	Review	The order of a reviewed essay for one reviewer: 1st, 2nd, 3rd, and $\geq$ 4th
Review revision	Review	Number of times that the review was submitted; categorized to two levels: no revision (i.e., submitted just once) and revision (i.e., submitted more than once)
Comment length by word	Review	The number of words for a comment
Average sentence length	Review	The average number of words in a sentence for a comment

Table 2
Descriptive statistics of predictors related to review timing, all the same across reviewing dimensions.

Measure	Relative Frequency
Deadline proximity	Late submission: 5% Last-minute submission: 5% Early submission: 90%
Review duration	Speeded: 4% Normal: 96%
Order of review	1st: 25% 2nd: 26% 3rd: 23%≥ 4th: 26%
Review revision	No revision: 84% Revision: 16%

2016); Gao et al., 2019), as the document quality is most relevant to the current assignment. It was calculated using median peer ratings; this metric could be calculated for all the 620 studied reviewers, not just the smaller subset for whom experts evaluated their essays. The correlation between median peer ratings and mean expert ratings of the 293 essays was 0.63, 0.69 and 0.48 for composite ratings, argument-related ratings and language-related ratings, establishing a similarly high level of validity evidence as reported in previous research (Li et al., 2016). Median instead of mean peer ratings were used to reduce the influence of extreme peer ratings, and median peer rating was previously reported to be an accurate measure of quality (Piech et al., 2013). However, using a mean of peer ratings produced almost identical results.

By contrast, essay quality (as a predictor) was measured using an average score of median peer ratings (of four peer raters on average) and mean expert ratings (of two experts) for two reasons. First, this average score involved more raters than using only peer ratings or expert ratings, which would presumably result in more reliable estimation of essay quality. Second, the dependent variable, the errors, was essentially an adjusted discrepancy measure between expert ratings and peer ratings (i.e., expert – peer); using only peer ratings or only expert ratings as essay quality measure resulted in an artificial relation with the dependent variable. Note however

Table 3
Descriptive statistics of predictors specific to each reviewing dimension.

Measure	Reviewing Dimensions		
	Composite	Argument	Language
Review error	Accurate: 69.1% Severe: 15.2% Lenient: 15.6%	Accurate: 63.4% Severe: 15.5% Lenient: 20.6%	Accurate: 64.3% Severe: 18.9% Lenient: 16.9%
Essay quality	4.96 (0.76)	4.82 (0.90)	5.10 (0.72)
Reviewer ability	5.39 (0.90)	5.22 (1.00)	5.57 (0.90)
Review disagreement	0.73 (0.36)	0.82 (0.39)	0.81 (0.40)
Comment length by word	279.1 (182.4)	172.4 (119.2)	107.1 (76.8)
Average sentence length	18.5 (5.8)	19.4 (6.6)	17.4 (6.5)

Note. Numbers stand for Mean (SD).

that analyses using either median peer ratings or mean expert ratings as the essay quality predictor produced similar effects for all other predictors.

Review difficulty was indirectly measured by reviewer disagreement – disagreement of reviewers on one essay, which is measured by the standard deviation of peer ratings on one essay.

3.6. Data analysis procedures

We used standardized coefficients for all continuous predictors to simplify the interpretation, including essay quality, reviewer ability, review disagreement, comment length by word, and average sentence length, while using non-standardized coefficients for categorical predictors. Frequency percentages of categorical measures and means and standard deviations of continuous measures are presented in Tables 2 and 3.

A series of two-level logistic regressions were conducted using the function "glmer" from the R package "lme4" (Bates, Mächler, Bolker, & Walker, 2015) to examine the relation between predictors and each review error type, using *Accurate* as the reference category. In the analyses, individualized binary logistic regressions were conducted separately for the two types of errors with a random essay-level intercept in the model (Becg & Gray, 1984). We did not include reviewer-level variation in a separate level given that there were a large number of reviewers (N = 620) involved and around 45% (N = 278 out of 620 reviewers) of the reviewers in the analyzed dataset only provided one review. Therefore, review-level predictors had a large overlap with reviewing process, and they were collapsed within the statistical model into a reviewing-process level. Level 1 (the reviewing-process level) is nested in level 2 (the essay level) of the regression model because one essay received approximately four peer reviews. The relationships between the different levels in the regression model are illustrated in Fig. 3. We combined data from the ten schools as a whole and did not model school as a separate level here. There were too few essays per school to separately test the models in each school's data alone. A Title I vs. non-Title I school predictor was not significant; further the same pattern of results was found when the models were run separately for Title I vs. non-Title I schools. The mathematical formula for the logit model is presented next.

Level 1 (reviewing-process-level):

$$\log\left(\frac{p(error_{ij})}{p(accurate_{ij})}\right) = \pi_{0j} + \sum_{p=1}^{P} \pi_p x_{pij} + \varepsilon_{ij}$$

Level 2 (essay-level):

$$\pi_{0j} = \beta_0 + \sum_{k=1}^{K} \beta_k w_{kj} + u_{0j}$$

in which terms were defined as follows:

p(accurate$_{ij}$): probability of review provided by reviewer i to essay j being accurate;

p(error$_{ij}$): probability of review provided by reviewer i to essay j being either severe or lenient as appropriate;

π_{0j}: the random intercept that varies across different essays (i.e., writers);

π_p: the level-1 coefficient for the pth predictor;

x$_{pij}$: the pth predictor at level-1 with the subscript *ij* referring to an essay-reviewer pair; the reviewer-level predictor Reviewer Ability was included at level-1;

β_0: the level-2 intercept;

β_k: the level-2 coefficient for the kth predictor;

w_{kj}: the kth predictor at level-2; and.

ε_{ij} and u_{0j}: level-1 and level-2 random errors.

This method is considered to be conservative since only a subset of the data is used for each individualized analysis (Becg & Gray,

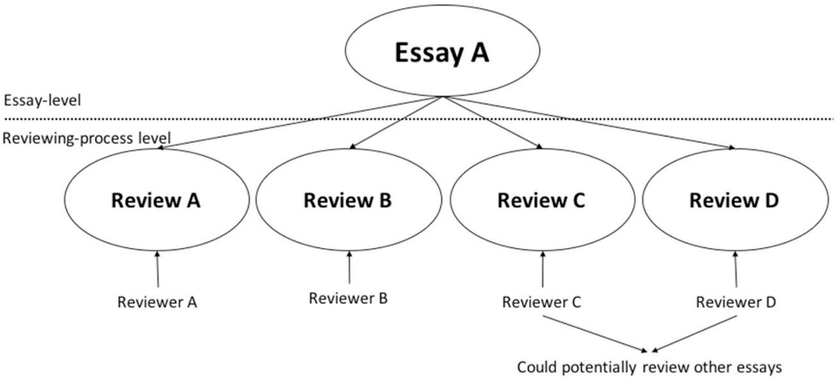

Fig. 3. Nesting structure of the peer review data in the regression models.

1984). We used the sample with accurate reviews and severe errors for the binary logistic regression model predicting severe errors; similarly, we used the sample with accurate reviews and lenient errors to predict lenient errors. We used a $p < 0.05$ statistical significance threshold. In addition, we also used Tjur's Coefficient of Discrimination (Tjur's D) as an R^2 measure to evaluate model performance (Tjur, 2009).

4. Results and discussion

4.1. *What characteristics predict web-based peer review errors?*

Significant Predictors of Overall Errors. Table 4 presents the correlations among the predictors. Most correlations between the predictors were very small. There were only a couple of exceptions: review disagreement has a small negative correlation with essay quality; comment length has a small positive correlation with reviewer ability, a small negative correlation with speeded review, and a medium correlation with average sentence length. Variance inflation factor (VIF) was also checked to ensure there were not multicollinearity issues. Focusing first on composite ratings (i.e., averaged across dimensions), five variables were statistically significant predictors of review errors — four for severe errors and two for lenient errors. Fig. 4 presents the odds ratios with 95% confidence intervals for these predictors. An odds ratio larger than one means that error is more likely as the predictor value increases, and a ratio less than one means the error is less likely as the predictor value increases. In addition, odds ratios were plotted using a log scaled x-axis of Fig. 4 in order to maintain symmetry between odds ratios greater and smaller than one. Interestingly, one variable (i.e., comment length) predicted both severe and lenient errors in opposite directions, three variables (i.e., review disagreement, reviewer ability, and average sentence length) only predicted severe errors, and one variable (i.e., essay quality) only predicted lenient errors.

Comment length has opposite relationships to each error type with a similar magnitude: longer comments were more likely to be in the severe category and less likely to be in the lenient category. In particular, with every standard deviation increase in comment length, the odds for the essay being severe rather than accurate (i.e., probability of being severe/probability of being accurate) increased by 0.55 times (*odds ratio* = 1.55, $p < 0.01$), but decreased by 0.79 times for the odds of being lenient rather than accurate (*odds ratio* = 0.56, $p < 0.01$). Therefore, longer comments were associated with more severe reviews but less lenient reviews. Of course, the causality is ambiguous and perhaps bidirectional: when students wrote detailed accounts of problems, they might exaggerate the severity, and/or if they were offended by errors they might feel compelled to write longer comments.

Review disagreement predicted severe errors. For each standard deviation increase in review disagreement, the odds of the overall essay rating being severe rather than accurate increased by 1.15 times (*odds ratio* = 2.15, $p < 0.01$). Where there is high disagreement, the review was more likely to be severe than accurate. Perhaps some students had different interpretation of requirements and thus

Table 4
Correlations between predictors.

	1	2	3	4	5	6	7	8	9
1. Essay Quality									
2. Reviewer ability	0.09** 0.08* 0.07*								
3. Review disagreement	−0.28** −0.22** −0.34**	−0.05 −0.09** −0.01							
4. Late submission	0.04 0.04 0.03	−0.02 0.01 −0.04	−0.03 −0.02 −0.01						
5. Last-min submission	−0.07* −0.07* −0.06	−0.06 −0.06 −0.06	0.02 0.04 0.00	−0.05					
6. Speeded review	0.03 0.03 0.02	−0.03 −0.02 −0.03	0.04 0.04 0.01	0.01	0.00				
7. Order of review	−0.04 −0.03 −0.04	0.02 0.02 0.02	0.03 0.02 0.04	0.13**	0.22**	0.12**			
8. Review revision	0.01 0.00 0.02	0.02 0.00 0.03	0.01 0.03 0.01	0.00	−0.05	−0.01	−0.06*		
9. Comment length	0.02 0.02 0.01	0.22** 0.21** 0.21**	−0.01 −0.03 0.01	−0.07*	−0.02	−0.20**	−0.05	0.06*	
10. Ave. sentence length	0.01 0.01 0.01	0.17** 0.17** 0.15**	−0.06* −0.08** −0.01	−0.07*	0.03	−0.19**	−0.05	0.03	0.52*

Note. *$p < 0.05$; **$p < 0.01$.
For the cells with three numbers, first row: correlation for composite scale; second row: correlation for argument dimension; third row: correlation for language dimension.

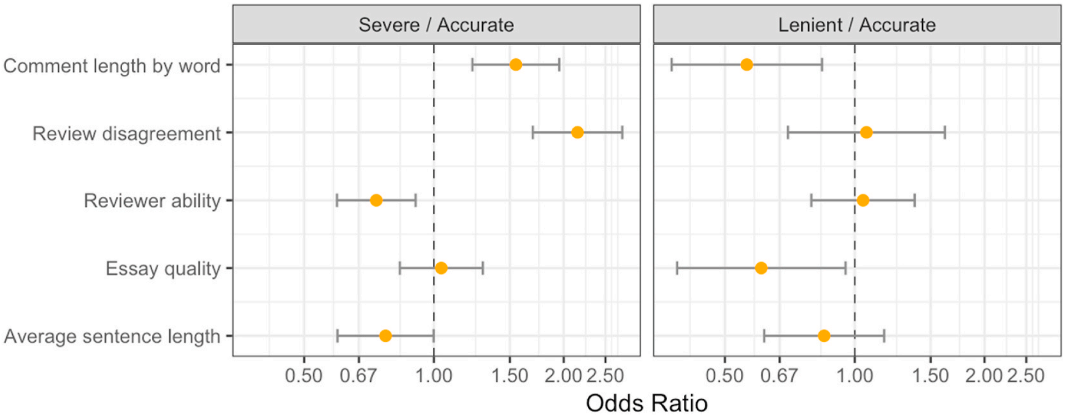

Fig. 4. Odds ratios and 95% confidence intervals for significant predictors of severity (left) and leniency (right). The x-axis is distorted to log scaling.

gave harsh ratings based on their unique understandings of the rubrics.

Another essay-level variable, essay quality was associated with lenient errors. Higher-quality essays were less likely to be in the lenient category. Specifically, for a standard deviation increase in essay quality, the odds for the essay being in lenient category (as compared to accurate category) decreased by 0.64 times (*odds ratio* = 0.61, $p < 0.05$). It may reflect a ceiling effect in which higher quality essays were too close to ceiling on the 7-point rubrics to produce lenient errors.

Reviewer ability was only predictive of severe errors, with no relationship to lenient errors. In particular, lower ability reviewers were more likely to produce severe ratings, with every standard deviation decrease in ability being associated with a 1.35 times in the odds of giving a severe rating than an accurate rating (*odds ratio* = 0.74, $p < 0.01$). Interestingly, this pattern is different from prior research finding that student reviewers with higher writing competence tend to provide more critical feedback in evaluating others' work (Strijbos, Narciss, & Dünnebier, 2010). However, review criticism and error are different concepts in that a critical comment is not necessarily error-prone. An explanation of the current result may be that weaker reviewers sometimes miscategorized some well-developed arguments or explanations as incorrect due to their lack of ability in understanding the arguments or explanations. This pattern may be related to the Dunning-Kruger effect — a metacognitive deficit of the unskilled in recognizing their own incompetence (Kruger & Dunning, 1999) and thus likely also in recognizing strengths/weaknesses in others (Huang, 2013).

Finally, average sentence length was predictive of severe error. In particular, longer sentences in comments were associated with less severe errors with one standard deviation increase of average sentence length by 1.30 times less likely in severe errors (*odds ratio* = 0.77, $p < 0.05$). The comments with more complex sentences, which were possibly more careful reviews, were slightly related to less severity and more accuracy. In contrast to the results involving comment length, average sentence length was associated with fewer errors. As a reminder, comment length and average sentence length are two correlated but different measures: Comment length is the length of the whole comment, and average sentence length is an indicator of sentence complexity. When both predictors were included in the model, average sentence length contributed to explaining review errors beyond overall comment length.

Tjur' D was 14.7% and 40.9% for the severe error model and the lenient error model respectively, indicating good explanative power of the models. Tjur' D is a similar measure as R^2 in multiple linear regression. Approximated by the Cohen's standard for Pearson's r (Cohen, 1992), a Tjur' D of 14.7% is between a medium and large effect size.

Nonsignificant Predictors of Overall Errors. Deadline proximity, review duration, review revision, and order of review were not significant predictors of either kind of review error in the composite level after controlling for the effects noted in the earlier analyses (see Appendix B for details).

Errors in Specific Writing Dimension Ratings. We further examined review errors in each of the five different dimensions. The patterns of results were generally similar to that at the composite level, however with some variation across argument-related vs. language-related dimensions. Therefore, we present two higher-level dimensions derived from the five dimensions: 1) argument-related (including rhetorical strategies, evidence for claims, and explaining evidence) and 2) language-related (including organization and control of language and conventions).

Some of the predictors showed larger effects with the argument dimension than with language dimension. For example, reviewer ability was only statistically significant for predicting severity in the argument dimension (*odds ratio* = 0.80, $p < 0.01$), but not in the language dimension. In addition, average sentence length was only significant for predicting severity in the argument dimension (*odds ratio* = 0.73, $p < 0.05$), but not in the language dimension. The effect of comment length on severity was also stronger in the argument dimension (*odds ratio* = 1.52, $p < 0.01$) than in the language dimension (*odds ratio* = 1.23, $p < 0.05$).

However, some of the predictors showed larger effects for the language dimension. For example, essay quality was only predictive of errors in language dimension (lenient errors: *odds ratio* = 0.42, $p < 0.01$), but not of errors in the argument dimension. That is, higher quality essays involved fewer lenient errors, but only in the language dimension. Detailed results can be found in Appendix B.

4.2. Are the same characteristics predictive across lenient vs. severe errors?

Looking across the results, the only factor that predicted both lenient and severe errors was comment length. However, it actually showed opposite relationships for the two types of errors: longer comments were positively associated with severe errors, but negatively associated with lenient errors. Review disagreement, reviewer ability and average sentence length only predicted severe errors while essay quality was only associated with lenient errors. These patterns indicate that the prediction of different types of errors are specific, which highlights the importance of considering severe vs. lenient errors separately.

4.3. Do all three levels of peer review predict errors?

Looking across the results (also see Fig. 5), there were factors from each level of the theoretical framework (i.e., connecting to each aspect of the dyadic nature of peer review) that predicted review errors. To further test this point, we also entered the predictors in three steps, following the framework sequentially by first entering the reviewer-level predictor, followed by essay-level predictors and reviewing-process-level predictors. The purpose of this stepwise procedure was to evaluate whether adding predictors from all levels was necessary. The χ^2 significance test results showed that adding each level improved the model fit significantly, which provides validity evidence for the three-level framework.

5. General discussion

5.1. Revisiting the framework of review errors

Most prior studies of overall reliability and validity of peer reviews have found acceptable levels of both reliability and validity in general. The present study went beyond this general level by separately examining the two types of review errors (i.e., leniency and severity) and possible factors that related to these two types of review errors so that more can be done to address errors that inevitably come for some students, even when overall reliability and validity is acceptable. Five characteristics out of the list of all predictors were found to be significant predictors, including characteristics at all three levels of the framework, with one predicting both types of errors and four only predicting one type of error. Fig. 5 presents the updated framework, which shows only the significant characteristics from the current study and the ways in which they were connected to each error type.

It is important to note four key methodological strengths of the current study. First, the current study more precisely measured errors through multiple trained expert raters, rather than relying on only a single expert or Olympic average across many peer raters (de Alfaro & Shavlovsky, 2016) as the gold standard. Second, it separately investigated two different types of errors, revealing asymmetries in what predicted each error type. Third, it included a broader range of characteristics, which make the study more rigorous (by including more controls) and more comprehensive. Fourth, the current study addressed the nested factors of reviewing process within essay (or author) using a multilevel modeling approach. The significant predictors found in this study at each of the three proposed levels (i.e., reviewer level, essay level, and reviewing-process level) highlight the value of studying peer review validity by including characteristics from different levels.

Essay-level significant predictors. Results of the binary logistic regressions showed that review disagreement was a strong predictor of severe errors in a positive way, consistent with the literature that ambiguous essays are more difficult to evaluate and thus result in more errors (Nguyen & Litman, 2013b). Although review disagreement as a measure of review difficulty was indirect and may also reflect higher levels of review noise, the results suggest that further investigation of related essay-level characteristics would be fruitful (e.g., essay characteristics that make it likely difficult to grade). Gao et al. (2019) found that certain kinds of writing issues appeared to be beyond the zone of proximal development for peer reviewers (i.e., never detected), whereas other kinds of writing issues were consistently detected by the peers.

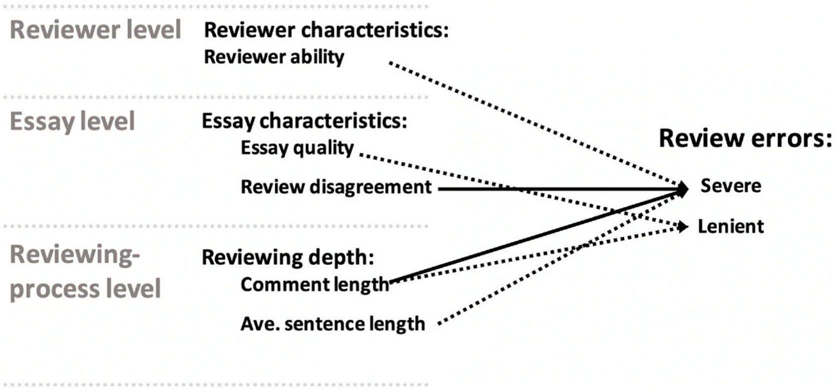

Fig. 5. Updated framework of significant reviewer, essay, and reviewing-process level predictors of severe and lenient review error types. Dashed lines refer to negative coefficients while solid lines refer to positive coefficients.

Another essay-level characteristic, essay quality, showed that higher quality essays were less likely to be rated leniently, and therefore more accurately. This essay quality effect is not consistent with either the "central tendency" effect (i.e., higher quality essays being rated lower than their actual quality and lower quality essays being rated higher than their actual quality) found in a traditional classroom (Matsuno, 2009) or the exaggeration effect (i.e., higher quality essays being rated even higher than their actual quality and lower quality essays being rated even lower than their actual quality) found in a MOOC study (Piech et al., 2013). One possible reason is that we have removed the general leniency tendency of peer reviewers as an adjustment of errors in the current study, so that the essay quality effect was more reflective of the actual effect. This effect, though, seems to be consistent with the positive relation between peer solution quality and peer-feedback accuracy in geometry proofs (Alqassab, Strijbos, & Ufer, 2018). This effect again suggests that looking into essay-level characteristics is important.

Reviewer-level significant predictors. Reviewer ability was found to be negatively related to severe errors, which is consistent with the intuition that lower-ability reviewers are most different from expert reviewers. A similar pattern was found in a MOOC — the weakest students had lower levels of accuracy in their peer reviews (de Alfaro & Shavlovsky, 2016). However, by distinguishing different types of errors, our results showed that the effect of reviewer ability was only on severe errors, indicating lower-ability reviewers tend to be more severe, which is somewhat unexpected. A possible explanation is that weaker reviewers may have been more likely to miscategorize some arguments or explanations as incorrect. Nevertheless, the reviewer ability effect is not very strong (*odds ratio* = 0.74), which is also consistent with the existing literature (Patchan, Hawk, Stevens, & Schunn, 2013, Patchan and Schunn, 2016; Matsuno, 2009).

Interestingly, the reviewer ability effect was found to be more similar to that previously found in a MOOC context than to those previously in a traditional higher education context. Despite the large difference between a typical MOOC and the current study context, one similarity between the current study and the MOOC is the diversity of essay quality as the current study included both high-performing and low-performing high schools. However, since students only reviewed within a school, students were exposed to less diversity in essay quality than in a MOOC. The mixed results from different studies also highlights the importance of situating peer review within a specific context. It will be important to create new models using the current framework for predicting review error in substantially different contexts.

Review-processing-level significant predictors. For the review-process characteristic comment length, overall longer comments were more likely to be related to severe errors and less likely to be lenient errors. By contrast, another review-process characteristic, the average sentence length within a comment was only significant in predicting severe errors in the current study context. The results of the two related but conceptually distinct review-depth characteristics provide further evidence of the important relationship between written comments and numerical ratings in peer review; Patchan et al. (2017) found that giving incentives for higher quality comments also improved the reliability of numerical ratings.

Overall, the current study presented a theoretically-grounded but easily operationalized framework for studying peer review errors, and this three-level framework can be used to study which factors will significantly predict review errors in various contexts. However, it is also important to acknowledge that other characteristics within each of the three levels could also prove to be more important. The current investigation focused on factors that are easily measurable in web-based peer review. But more complex processes could be built that leverages information on past performance of each reviewer, or automatically examines topic complexity or novelty in the essay (Li, Chi, Li, Ouyang, & Fu, 2006), the negative or positive focus of comments (Liu, 2012), or the use of specific constructive advice in comments (Nguyen & Litman, 2013a).

5.2. *Implications for practice*

Results of the current study could be applied to different peer review or general assessment settings by helping to detect errors or by directing new kinds of support for promoting assessment accuracy and fairness. For example, the particular characteristics identified in this study could be the foundation for development of new tools that could automatically flag reviews likely to be inaccurate. A recent study has reported a machine learning method in automatic detection of inconsistency between numerical scores and textual feedback (Rico-Juan et al., 2019). The results of this study provide insights in further development of new tools that leverage automatic classification based on the features identified in this study.

For example, using the reviewer and author characteristics, balanced pools of specific reviewers as well as total numbers of reviewers can be determined to efficiently allocate resources as well as minimize the chance that certain documents will be mis-graded. The essay quality and reviewer ability effects found in the current study suggest that it is important to have a balanced distribution of peer reviewers to essays to make sure each writer receives feedback from diverse ability reviewers. Ensuring that all students receive triangulated reviews could potentially reduce the overall impact of extreme errors.

Using all three levels, scores could be automatically adjusted by down-weighting or removing reviews that are likely to be inaccurate. However, the accuracy of the prediction would likely need to be high to use such a method, and that level of accuracy was not met in the current dataset, which needs further development. Alternatively, reviews with high likelihood of error could be flagged for TA or instructor oversight. This approach does not require as high a level of model accuracy.

These effects might also be the foundation of new facilitation to peer reviewers. For example, the situations most likely to produce review errors could launch additional scaffolds to guide peer reviewers. For example, there could be reviewer-level interventions implemented during the review process such as automated motivational messages sent to reviewers to motivate them to produce thoughtful comments along with ratings in the case of overly short reviews that were predictive of reviewer errors.

5.3. *Limitations and future research*

Our study has several limitations. First, some of the measures were indirect. For example, review disagreement was an indirect measure of review difficulty. The strong observed effect of review disagreement argues for follow-up studies that include other essay characteristics measuring review difficulty more directly (e.g., whether a specific essay includes writing approaches that the rating rubrics do not cover). Reviewer ability measured by their writing performance is also an indirect measure. Reviewing ability in terms of reviewers' understanding of rating rubrics or ability in detecting misunderstandings should be further investigated. Review duration was indirectly measured as the time duration between requesting reviews and submitting reviews, and we did not have the information on the actual time a student reviewer worked on a review task.

Second, while we have included a broad range of variables in the model, the list is not exhaustive. Inclusion of other variables in the model might produce even better prediction of review errors. For example, the textual characteristics from the reviewed essays may be useful information to include. In addition, the strong predictive power of comment length also indicates that studying other deep characteristics of textual comments may be worthwhile. However, prediction models that leverage such features might be less generalizable across contexts/assignments.

Third, the current study focused on the first draft of the essay writing. It would be interesting to study whether the results found in the current study also generalize to essay revision rounds or later assignments. The relative balance of errors of severity and leniency may change over time.

Fourth, the current study was conducted in a secondary school peer review context in an AP writing course with a set of well-designed reviewing prompts, which echoes back to the positioning the current study in the micro-level lens in which the macro-level context is predominantly fixed (e.g., a consistently structured high school writing course, a common writing assignment, following shared procedures, shared rubrics, and a shared web-based peer review system). Although some previous studies have suggested that subject discipline or details/forms of the peer review tasks (e.g., oral presentation vs. writing) did not impact the validity of peer review results (Li et al., 2016) or its effect on learning (Sanchez et al., 2017), we have not tested the model examined in the current study under different contexts to examine the generalizability of the results. We have avoided using features that are specific to the particular writing assignment (e.g., predictors for specific dimensions) or based on the content of the submissions themselves because those will likely be more difficult to generalize across contexts. Further, we have used an assignment that involves many different reviewing dimensions and collected data from a purposely diverse set of students and school contexts, spanning from high level to low level issues to improve the generalizability of the findings. However, future studies using the framework under different peer review contexts are needed to fully test the generalizability. This will be especially important for achieving the practical goal of building tools that identify peer errors and to provide supportive feedback to both students and instructors on how to improve review quality.

Lastly, in addition to examining model generalizability under different contexts, studies incorporating both micro-level and macro-level variables at once could be fruitful, especially since the micro level and macro level are not unrelated parallel levels, but rather reciprocally interconnected. For example, an effective rater training program with examples and practices, which is at the macro-level, could shifts the distribution of individual reviewer abilities, resulting no last minute reviews or less within-class variation in reviewer ability, which are micro-level variables. Nevertheless, an empirical study with such a wide range of levels and complicated interconnected variables would be challenging given the likely need for massive amount of expert data under different peer review implementation contexts (multiple experts ratings supplied for a sufficiently large number documents within each context to support complex statistical models). However, with the availability of web-based peer review systems being used by a large number of instructors and institutions, this kind of empirical study may become possible in the future.

Author credit statement

Yao Xiong: Conceptualization, Data curation, Formal analysis, Visualization, original drafting, reviewing and editing; Christian D. Schunn: data acquisition, Conceptualization, Formal analysis, Visualization, original drafting, reviewing and editing.

Acknowledgement

The current project was partially funded by grant R305A120370 from the Institute of Education Sciences.

Appendix A

Rating Rubrics.

Rating	7	5	3	1
Rhetorical strategies	The author analyses multiple, subtle rhetorical strategies that	The author analyses three or more obvious rhetorical	The author analyses only 1–2 obvious rhetorical strategies that	The author didn't write about Kelley's rhetorical strategies

(*continued on next page*)

(*continued*)

Rating	7	5	3	1
	Kelley uses accurately (such as appeal to a common cause, evoking nostalgia, or other sophisticated strategies).	strategies that Kelley uses (such as using rhetorical questions, anecdotes, or other obvious strategies).	Kelley uses (such as rhetorical questions) or misunderstands Kelly's strategies.	(instead discussed a different topic, connected to personal experience, or just summarized Kelly's piece).
Evidence for claims	Every claim has accurate evidence for all important aspects of the claim. Most evidence is conveyed through direct quotes.	Every claim has evidence, but some of the evidence is not accurate or not complete. Some evidence	Every claim has evidence, but some of the evidence is not accurate or not complete. Some evidence	No evidence is provided for any of the claims.
Explaining evidence	Explanations of all the evidence provided are thorough, logical and connected to the essay's thesis.	Explanations are sufficient, but not always thorough, logical, and clearly connected to the essay's thesis.	Explanations are simplistic, sometimes absent, or not clearly connected to the essay's thesis.	Explanations are missing or unrelated to the prompt (such as based in personal experience).
Organization	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.	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.	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.	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	Mature, sophisticated prose style, using specific academic terminology (such as pathos and ethos) and control of language.	Clear prose style with few lapses in academic word choice.	The prose generally conveys the writer's ideas but is inconsistent in controlling the elements of effective writing, such as academic word choice.	Simplistic style and vocabulary.
Conventions	The paper follows the conventions of Standard Written English very well with very few or no errors.	The paper mostly follows the conventions of Standard Written English, but has about 1–2 errors per paragraph. The errors don't interfere with your understanding the writer's ideas.	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.	In many sentences, the paper does not follow the conventions of Standard Written English. The errors make it very difficult to understand the writer's ideas in many places.

Appendix B

Results for Composite and Dimensional Errors.

Dimension	Error	Predictor	Odds ratio	*z*-value	*p*-value
Composite	Severe	Essay quality	1.04	0.35	0.72
	Tjur's D: 14.7%	Reviewer ability	0.74	−2.87	0.00
		Review disagreement	2.15	6.29	0.00
		Late submission	0.56	−0.84	0.40
		Last-min submission	0.97	−0.07	0.94
		Speeded review	2.44	1.58	0.11
		Order of review	1.06	0.61	0.54
		Review revision	0.80	−0.79	0.43
		Comment length	1.55	3.71	0.00
		Ave. sentence length	0.77	−1.97	0.05
	Lenient	Essay quality	0.61	−2.18	0.03
	Tjur's D: 40.9%	Reviewer ability	1.05	0.32	0.75
		Review disagreement	1.06	0.29	0.77
		Late submission	0.40	−1.38	0.17
		Last-min submission	1.36	0.55	0.58
		Speeded review	2.55	1.33	0.18
		Order of review	1.12	0.92	0.36
		Review revision	0.52	−1.74	0.08
		Comment length	0.56	−2.81	0.00
		Ave. sentence length	0.85	−1.00	0.32
Argument	Severe	Essay quality	1.02	0.13	0.90
	Tjur's D: 16.1%	Reviewer ability	0.80	−2.08	0.04
		Review disagreement	1.95	5.33	0.00
		Late submission	0.46	−1.14	0.26
		Last-min submission	1.01	0.01	0.99

(*continued on next page*)

(*continued*)

Dimension	Error	Predictor	Odds ratio	*z*-value	*p*-value
		Speeded review	1.10	0.16	0.87
		Order of review	1.12	1.22	0.22
		Review revision	0.94	−0.24	0.81
		Comment length	1.52	3.55	0.00
		Ave. sentence length	0.73	−2.37	0.02
	Lenient	Essay quality	0.85	−1.16	0.25
	Tjur's D: 23.5%	Reviewer ability	1.07	0.65	0.51
		Review disagreement	1.19	1.33	0.18
		Late submission	0.59	−1.03	0.30
		Last-min submission	1.36	0.69	0.49
		Speeded review	0.78	−0.44	0.66
		Order of review	1.19	1.84	0.07
		Review revision	0.62	−1.63	0.10
		Comment length	0.66	−2.86	0.00
		Ave. sentence length	0.93	−0.58	0.56
Language	Severe	Essay quality	0.94	−0.67	0.50
	Tjur's D: 9.8%	Reviewer ability	0.91	−1.05	0.29
		Review disagreement	1.86	5.80	0.00
		Late submission	0.66	−0.73	0.47
		Last-min submission	0.99	−0.03	0.98
		Speeded review	2.35	1.81	0.07
		Order of review	0.98	−0.27	0.79
		Review revision	0.69	−1.44	0.15
		Comment length	1.23	2.03	0.04
		Ave. sentence length	0.92	−0.76	0.45
	Lenient	Essay quality	0.42	−3.71	0.00
	Tjur's D: 42.4%	Reviewer ability	1.14	0.96	0.34
		Review disagreement	0.99	−0.06	0.95
		Late submission	0.56	−0.92	0.36
		Last-min submission	1.67	0.93	0.35
		Speeded review	1.09	0.11	0.91
		Order of review	1.15	1.18	0.24
		Review revision	0.60	−1.39	0.17
		Comment length	0.66	−2.28	0.02
		Ave. sentence length	0.93	−0.50	0.62

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2021.104146.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (Washington, DC).

de Alfaro, L., & Shavlovsky, M. (2016). Dynamics of peer grading: An empirical study. In *The 9th international conference on educational data mining* (pp. 62–69).

Alqassab, M., Strijbos, J.-W., & Ufer, S. (2018). The impact of peer solution quality on peer-feedback provision on geometry proofs: Evidence from eye-movement analysis. *Learning and Instruction, 58*, 182–192.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using "lme4. *Journal of Statistical Software, 67*(1), 1–48.

Becg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika, 71*(1), 11–18.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*(2), 185–200.

Chang, C.-C., Tseng, K.-H., Chou, P.-N., & Chen, Y.-H. (2011). Reliability and validity of web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education, 57*(1), 1306–1316.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338.

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.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Elizondo-Garcia, J., Schunn, C. D., & Gallardo, K. (2019). Quality of peer feedback in relation to instructional design: A comparative study in energy and sustainability MOOCs. *International Journal of Instruction, 12*(1), 1025–1040.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

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.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290.

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.

Huang, S. (2013). When peers are not peers and don't know it: The Dunning-Kruger effect and self-fulfilling prophecy in peer-review. *BioEssays, 35*(5), 414–416.

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101–110.

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.

Institute of Education Sciences. (2007). *National assessment of Title I final report: Summary of key findings*.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Langer, J. A., & Applebee, A. N. (1987). *How writing shapes thipking: A study of teaching and learning*. National Council of Teachers of English.

Li, X., Chi, J., Li, C., Ouyang, J., & Fu, B. (2006). Integrating topic modeling with word embeddings by mixtures of vMFs. In *International conference on computational linguistics: Technical papers* (pp. 151–160).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Liu, J., Guo, X., Gao, R., Fram, P., Ling, Y., Zhang, H., et al. (2019). Students' learning outcomes and peer rating accuracy in compulsory and voluntary online peer assessment. *Assessment & Evaluation in Higher Education, 44*(6), 835–847.

Li, H., Xiong, Y., Zang, X., Kornhaber, M., Lyu, Y., Chung, K. S., et al. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264.

Luo, H., Robinson, A. C., & Park, J.-Y. (2014). Peer grading in a MOOC: Reliability, validity, and perceived effects. *Journal of Asynchronous Learning Networks, 18*(2). Retrieved from onlinelearningconsortium.org/sites/default/files/429-2286-1-LE.pdf.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Nguyen, H. V., & Litman, D. J. (2013a). Identifying localization in peer reviews of argument diagrams. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (Vol. 7926). Berlin, Heidelberg: Springer.

Nguyen, H. V., & Litman, D. J. (2013b). Predicting low vs. high disparity between peer and expert ratings in peer reviews of physics lab reports. In *International conference on artificial intelligence in education* (pp. 687–691). Berlin, Heidelberg: Springer.

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.

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.

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.

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2017). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education, 43*(12), 2263–2278.

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

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th international conference on educational data mining (EDM 2013)*. Memphis, Tennessee.

Popham, W. J. (1989). *Modern education measurement: A practitioner's perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15.

Rico-Juan, J. R., Gallego, A.-J., & Calvo-Zaragoza, J. (2019). Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. *Computers & Education, 140*, 103609.

Sadler, P., & Good, E. (2006). The Impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31.

Saeed, M. A., & Ghazali, K. (2017). Asynchronous group review of EFL writing: Interactions and text revisions. *Language Learning & Technology, 21*(2), 200–226.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066.

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., & DeMartino, 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.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27*(3), 161–168.

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.

Suen, H. K. (2013). Role and current methods of peer assessment in massive open online courses (MOOCs). In *The first international workshop on advanced learning sciences (IWALS)*. Pennsylvania: University Park.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *International Review of Research in Open and Distance Learning, 15*(3). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1680/2904.

Tjur, T. (2009). Coefficients of determination in logistic regression models — a new proposal: The coefficient of discrimination. *The American Statistician, 63*(4), 366–372.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Tsivitanidou, O. E., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction, 21*(4), 506–519.

Wu, W. C. V., Petit, E., & Chen, C. H. (2015). EFL writing revision with blind expert and peer review using a CMC open forum. *Computer Assisted Language Learning, 28*(1), 58–80.

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