Computers in Human Behavior 127 (2022) 107056

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Do experiences of interactional inequality predict lower depth of future student participation in peer review?

Zheng Zong [a], Christian D. Schunn [b], Yanqing Wang [a,*]

[a] School of Management, Harbin Institute of Technology, Harbin, 150001, China
[b] Learning Research & Development Center, University of Pittsburgh, PA, 15260, USA

ARTICLE INFO

Keywords:
Peer review
Peer feedback
Interaction inequality
Depth of the participation
Game theory

ABSTRACT

Across disciplines and educational levels, peer feedback has emerged as a broadly useful pedagogical strategy. However, the value of peer feedback depends upon students being willing to provide each other substantial feedback. We develop a novel application of game theory to study whether students' experiences with interaction inequality in peer feedback predict lower participation in future peer feedback assignments. Two kinds of inequality are explored: inequality in amount and inequality in the helpfulness of received vs. providing feedback. We examine data from students ($N = 732$) enrolled in the three different courses (varying in discipline and level) using the same online peer review platform. We use advanced multiple regression models to test four theoretically-derived hypotheses regarding the overall effects of experienced inequality, three hypotheses involving contextual moderators of that effect. The results confirmed the hypothesized relationship of experiencing inequality in the number and quality of feedback provided vs. received with the subsequent depth of participation in peer reviewing. The results also generally confirmed the three predicted moderators of the relationship: stronger observed relationships with inequality experiences for weaker students, in later assignments, and in more advanced coursework. This study is unique in applying and extending game theory to computerized peer review and in its approach for studying the relationship of specific prior experiences with future peer review participation. The findings provide new practical insights into what is essential to manage in peer review processes.

1. Introduction

Peer review, defined as having peers within a class provide numerical assessments and feedback comments to one another, has long been identified as a productive way for providing timely and effective feedback (Topping, 1998). Peer review has been used for a wide range of student artifacts, such as essays (Cho & Schunn, 2007; Schunk & Zimmerman, 2007; Applebee & Langer, 2011), video presentations (Min, 2016), design projects (Grabe & Kaplan, 1996; Gatfield, 1999), and computer code (Ballantyne et al., 2002; Wang et al., 2012; Wang & Sun, 2018). Meta-analyses have established that peer assessments are generally reliable (Li et al., 2016) and that peer feedback is typically useful for learning (Double et al., 2019; Li et al., 2020) and improving motivational outcomes (Li et al., 2021).

However, peer review benefits depend upon students actively participating in the process (sometimes referred to as behavioral engagement), and students are sometimes reticent to participate. Some students hold a variety of negative attitudes towards peer assessment (Mangelsdorf, 1992; Liu & Carless, 2006; Kaufman & Schunn, 2011; Zou et al., 2018). As a result, students sometimes refuse to participate entirely or participate at minimal levels (Patchan et al., 2018; Zou et al., 2018; Elizondo-Garcia et al., 2019). Some approaches have been proposed to encourage strong participation, including grading incentives for high-quality participation (Patchan et al., 2018) or training approaches (Russell, 2004). Prior research on peer review has not examined the role that prior experiences students with peer review have in shaping their future participation levels in peer review.

Building upon game theory research, we explore a set of hypotheses based upon the general notion that the reciprocal nature of peer review unfolding across rounds of reviewing. In particular, we predict students will reduce their participation across reviewing rounds when they experience inequality in reviewing in the prior round. Here inequality is defined as differences between the number of comments or helpfulness of comments provided to their peers vs. received from peers. The

* Corresponding author. School of Management, Harbin Institute of Technology Address: 13 Fayuan St., Nangang District, Harbin, 150001, China.
E-mail address: yanqing@hit.edu.cn (Y. Wang).

https://doi.org/10.1016/j.chb.2021.107056
Received 22 December 2020; Received in revised form 10 October 2021; Accepted 12 October 2021
Available online 20 October 2021

hypotheses are tested in three university courses in which students experience multiple assignments across the semester.

2. Theoretical background

Game theory was originally developed to explain patterns in individuals' decisions to cooperate or compete with others in a context (Binmore, 1994), and it has been extensively applied in the social sciences, computer science, and philosophy (e.g., Balliet et al., 2011; Conybeare, 1984; Corfman & Lehmann, 1994; Tangpong et al., 2010). A classic game theory example involves the prisoner's dilemma, in which two members of a criminal organization are asked to testify against one another. The payoff matrix is such that 1) the best outcome for each is obtained if they testify against the other person, but the other person does not testify against them; 2) the worst outcome is the converse; 3) the second-best outcome is if both stay silent, and 4) the second-worst outcome is if both testify. This problem is interesting because from an individual perspective, testifying appears to be the rationale choice and is what is sometimes observed as typical behavior (Hayashi et al., 1999; Kiyonari et al., 2000; Xi et al., 2013), but from the group perspective, the best average outcome occurs when both stay silent and this result is also sometimes observed (Anbarci & Feltovich, 2013; McNamara & Leimar, 2010). Empirical studies of repeated prisoner's dilemma problems, in which the same decision is repeated over rounds, have found that people react over time to how others have acted towards them in the prior round. In particular, individuals will act selfishly when the peer acted selfishly in the prior round or will cooperate if the peer acted that way in the prior round (Press & Dyson, 2012; Stewart & Plotkin, 2013).

Decisions in a number of real-world contexts (e.g., paying taxes, recycling) have this general structure (i.e., groups of people individually making the same decision repeatedly over time) and this kind of a payoff matrix (i.e., collaborating with others produces the best average outcome but an individual does best if they act selfishly). We argue that a similar game theory framework can also be usefully applied to the case of peer feedback, especially, how much feedback students will contribute and how they will react to how much feedback their peers contributed across rounds of reviewing. More generally, peer feedback is a kind of social information exchange, and equality in it has been previously examined within firms and online professional communities (e.g., Lin et al., 2009; Kleine et al., 2016; Lacey et al., 2017; Hayibor, 2017). In addition, interactive equality in general can involve both quantitative and qualitative aspects (Onwuegbuzie, 2012; Yilmaz, 2013).

Turning to the details of game theory applied to peer feedback, peer feedback is the intangible knowledge resource that students use to learn and improve their academic performance. Some prior work has argued that how students grade each other (Wu et al., 2015) or whether reviewees accept the feedback they receive could be shaped by a game theoretic analysis (Klein, 2018). However, no prior empirical work has examined how much feedback reviewers provide, which is the larger learning opportunity and also is necessary to having feedback that a student can accept. How much students gain from feedback depends upon the character of the feedback. Providing useful feedback involves effort, and there is some level of competition among students (e.g., for scholarships, for positions in more advanced degrees), and so receiving useful feedback from peers without providing useful feedback to peers could be perceived as the most advantageous outcome for the individual (Pandey & Chatterjee, 2016). However, if no useful feedback is provided by any peer, everyone experiences less learning. Thus, it could be argued that the overall best collective benefit within the class is when everyone provides useful feedback. In sum, the same conditions of prisoner's dilemma apply, and a similar prediction could be made across rounds of experience with peer feedback (across assignments) within a class: if peers provide less useful feedback, the student will also reduce the usefulness of the feedback the student provides in the next round, whereas if peers provide more useful feedback, the student will increase the usefulness of the feedback provided in the next round.

In the context of peer review, the simplest definition of quantitative inequality consists of the number of comments received, which can vary widely across reviewers (Patchan & Schunn, 2015). Consider the case of a student who gave each of the authors they reviewed 20 comments per document but only received ten comments from each of the students who reviewed their contribution. The student would likely feel the information exchange was unequal.

In addition to the number of comments, the comments' nature will likely matter, too. Sometimes a received peer comment is not perceived as helpful by the author, perhaps because the location of the problem was not clarified or the reason for the issue being a problem was not explained, or no possible solutions were offered (Nelson & Schunn, 2009; Patchan et al., 2016; Wu & Schunn, 2020). Indeed, without incentives for providing high-quality reviews, peers can provide review comments that consist of a short, generic praise comment (Patchan et al., 2018), like "nice job!". Thus, if a student gives each author ten long comments that are well explained and constructive but only receives ten short comments that are neither well explained nor constructive, the student would also be likely to feel the information exchange was unequal. Of course, not every constructive comment is seen as implementable, nor is every explanation seen as correct. Thus, the mere presence of explanations or suggestions for improvement does not directly define interactional equality. Instead, we argue that the perceived quality of the given and received comments (however, the students perceive comment quality) will drive the perception of interactional equality.

How might a student respond to inequality in the reviews regarding their peer review participation, the focus of the current research? Like reviewing experiences, the student's future participation in reviewing can vary in quantitative and qualitative terms. Quantitatively, the student can choose to vary the number of comments they provide. Qualitatively, the student could choose to give more superficial or more in-depth comments, which we will approximate by whether the student tends to give long comments.

2.1. Conceptual model and study hypotheses

Fig. 1 shows the overall conceptual model covering four hypotheses, beginning with a general hypothesis about the overall effect of experiencing inequality.

Hypothesis 1. Experiencing reviewing inequality (either quantitative or qualitative) in the prior peer review assignment will be associated with lower future reviewing participation.

We also have several hypotheses related to contextual moderators. Whether or not students do well in the assignment being reviewed may shape how they respond to reviewing inequality. For example, students

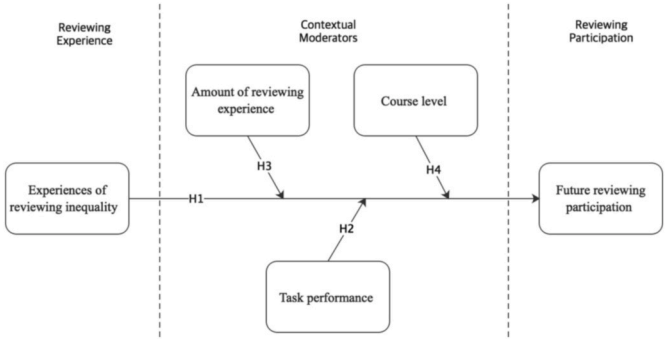

Fig. 1. The conceptual framework for the main relationship between prior experiences of inequality in reviewing with future participation in reviewing, with moderators of the amount of reviewing experience, course level, and own task performance.

with higher competency might be expected to provide more feedback because they have more relevant skills and knowledge (i.e., equality of effort means they should be providing more). Conversely, students with lower competency might feel they need more feedback (i.e., equality of value means they should receive more to reach the same end point). Both factors produce the same hypothesis:

Hypothesis 2. The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified when students do poorly in the course.

The total amount of experience in the discipline or specifically with peer review in the discipline may lead to norm-setting for what is expected in terms of participation in the peer review, making the inequality more salient. Thus, with a stronger norm (later assignments or later courses), students might have a stronger reaction to rating inequality.

Hypothesis 3. The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified with later peer review assignments.

Hypothesis 4. The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified in more advanced courses.

These four hypotheses are tested in three courses involving multiple peer review assignments. Because the courses all used a well-structured online peer review system, inequality in reviewing experiences was easily captured and quantified.

While the hypotheses are tested in the specific context of peer reviewing in coursework, these hypotheses should also apply to other forms of peer review (e.g., of conference papers, journal articles, research grants, and teaching) and knowledge information exchanges (e.g., *reddit* or *yahoo answers*), although with slight adaptions to those contexts (e.g., what is performance and what is more and less advanced 'courses'). The current study is also one of the first to systematically look at both qualitative and quantitative aspects of social information exchanges from a game theory perspective in any context. In particular, Hypothesis 1 examines whether quantitative and qualitative components are similarly important, which might be a novel contribution to game theory research. Both because of the replication crisis in social sciences (Maxwell et al., 2015) and the increasing need for theoretically-guided research to inform practice (Hughes, 2000), research that examines moderators of main theoretical predictions is becoming increasingly important: when are simple theoretical predictions observed (or stronger) and when are they not observed (or weaker)?

3. Methods

3.1. Participants

Three courses were selected for multiple reasons. First, they used the same online peer assessment system (Peerceptiv). Second, all had between 5 and 6 assignments, which enabled the study of change with experience (i.e., Hypothesis 3). Third, all were sufficiently large to provide sufficient statistical power to tease apart the influence of correlated factors using advanced statistical methods. To test the generality of the findings across disciplines, two of the courses were in Biology, and one was in Psychology. Finally, to test course level effects (i.e., Hypothesis 4), two of the courses were introductory survey courses (*Introductory Psychology* and *Introductory Biology*), and the other was an advanced course (*Advanced Biology*).

In total, there were 732 undergraduates across the three courses: *Introductory Psychology*–two different offerings of the same 2nd-level introductory psychology course (for majors and non-majors) at a large research-oriented public university in the Eastern part of the US; *Advanced Biology*–a for-majors writing-intensive biology course at a large research-oriented public university in the Western part of the US; and *Introductory Biology*–an introductory biology course (for non-majors) in a different large research-oriented public university in the Western part of the US. The data from the two offerings of the psychology course are treated as one course in the analyses because the assignments were nearly identical, and exploratory analyses found no differences in observed results within each offering. Table 1 shows the self-reported demographics details of students across the three courses.

3.2. Course setting

The *Introductory Psychology* and *Introductory Biology* courses had writing assignments aligned to different topics covered in the course. The reviewing rubrics involved examining some general writing elements like spelling, grammar, and other aspects of good writing, as well as clear and accurate descriptions of the selected scientific topics. The *Advanced Biology* course, by contrast, involved a sequence of writing assignments that supported a research project. Initially, students wrote about a research idea. Then they turned to summarize related prior research. Then they wrote about their planned experiment. Finally, they turned in a report that integrated all of the sections of a research report. The specific reviewing rubrics varied based on the writing assignment's nature and had a mixture of general writing quality and content-specific dimensions.

4. Materials

All students were required to submit assignments and complete peer reviews through a common online system, Peerceptiv (https://peerceptiv.com). Fig. 2 shows the main student reviewing interface. Students used this page to review their peers' submissions and enter open-ended feedback in the text boxes as guided by different reviewing prompts for each review dimension. Each assignment had separate reviewing dimensions (varying between 3 and 6 per assignment) created by the course instructor that directed reviewer attention to specific aspects of the writing assignment (e.g., style, writing conventions, organization, quality of explanations, the accuracy of content). Students had to submit at least one comment per reviewing dimension, but there was no minimum comment length. Each student had to review at least four peers' documents for each assignment. These ratings were used to produce grades for the student's submissions. All reviewing was double-blind—an author pseudonym identified author documents, and a number identified the reviewer. The system kept track of submitted documents, review ratings, and comments.

To incentivize higher review quality (and to measure reviewing inequality), authors had to rate the helpfulness of their received comments on a 1 to 5 scale, and optionally could provide an explanation for the rating (see Fig. 3 top). These ratings produced a reviewing quality grade. Reviewers could also see a mean helpfulness rating across all the comments they provided in a given review (see Fig. 3 bottom).

4.1. Measures

The measures used in this study were all derived from data automatically recorded by Peerceptiv (see Table 2 for definitions). The interactional component in repeated decision applications of game theory is fundamentally about individuals' decisions in response to the decisions of others from a prior round. The quantitative aspects of inequality in feedback are most directly about behavior counts (i.e., amount of feedback provided). The qualitative aspects are more inherently about perceptions (i.e., perceived usefulness), and Peerceptiv includes a measure of the perceived value of peer feedback as part of the system.

To calculate predictor and outcome variables, the data was organized by assignment, computing a value for each student in each

Table 1
For each course, the number of participating students, mean age, % female, and % reporting each race/ethnicity.

Course	# of Students	Mean age	% Female	Race/Ethnicity			
				% Asian	% Black	%Latinx	% White
Introductory Biology	98	19	82%	42%	3%	28%	28%
Advanced Biology	274	20	59%	69%	2%	14%	15%
Introductory Psychology	360	21	74%	8%	10%	0%	82%

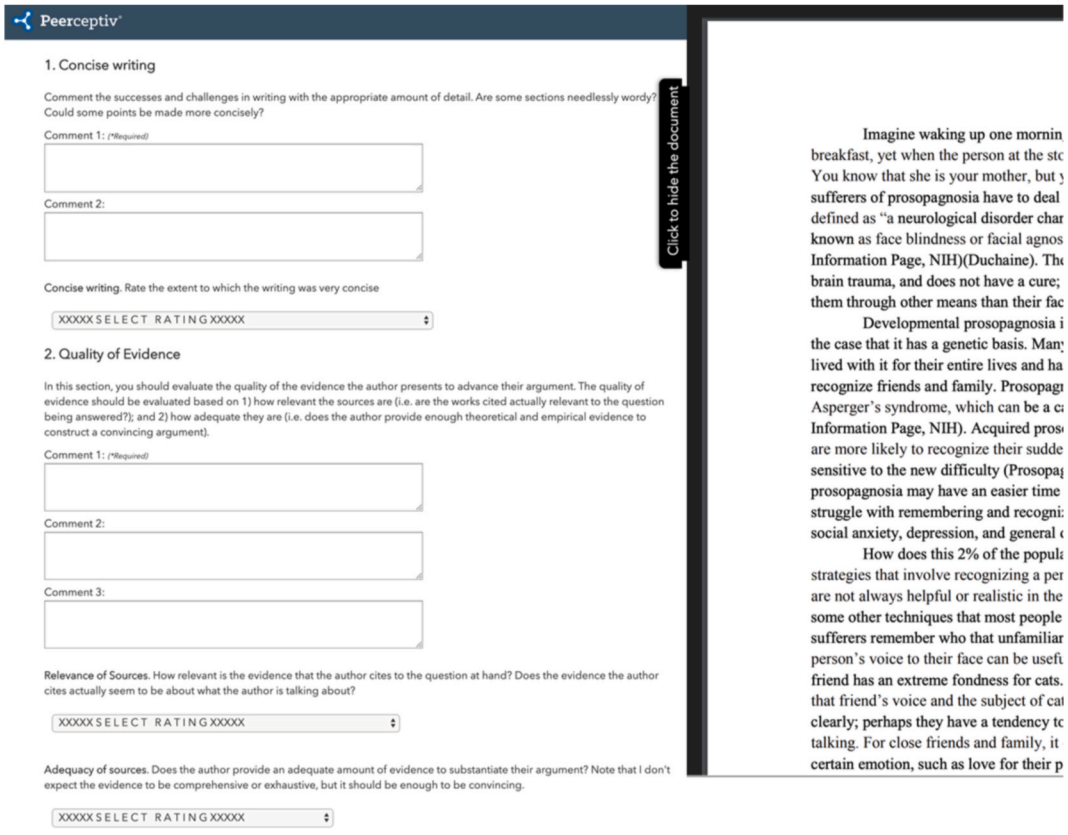

Fig. 2. The reviewing interface within *Peerceptiv* at the time of the study. Students can view the document (on the right) while completing the review (on the left). The review is divided into dimensions, with comments and ratings within each dimension. Students must enter at least one comment in the textboxes for each reviewing dimension.

assignment (e.g., # of long comments provided to peers on a given assignment). In this way, changes from one round to the next could be tested. For predictor values, only the first through second-to-last assignments in a course are relevant. For outcome values, only the second through last assignments in a course are relevant. The resulting *N* for analysis was 3202 across all three courses (732 students x 4–5 assignments). Table 3 lists the descriptive statistics for each variable within each course (Appendix A presents these descriptive statistics separately for each assignment within each course, see Table A1, Table A2, and Table A3 for details). Since two of the variables were used as both predictor and outcome and had different assignment number ranges in those roles, separate means and SDs are listed for the predictor and outcome roles.

4.1.1. Reviewing participation

The primary outcome is reviewing participation, which is measured in terms of long comments. Prior research on peer feedback (e.g., Nelson & Schunn, 2009; Patchan et al., 2016; Wu & Schunn, 2020) has found that short comments can rarely provide useful information to the author, potentially missing information about the problem's nature, where the problem occurs, suggestions for addressing the problem, and explanations for why it is important. For example, consider comments made by two different reviewers from one assignment in one of the courses:

In your second body paragraph, your topic sentence said that skin color determines which ethnicity you belong to. Unless I've misunderstood what you're trying to say, this is very incorrect. Skin color does NOT determine your ethnicity. Your ethnicity is determined by your ethnic ancestry/genes, not the color of your skin. For example, that would mean that all fair-skinned people are presumably Caucasian. I think you should revise your thoughts or clarify your ideas. When it comes to discussing how people evolved from their geographical location, this applies for the idea of body hair and how some people are hairier than others for body temp reasons. (107 words).

Although I didn't see any spelling mistakes, I did find a few problems regarding the sentence structure. (17 words).

The two comments vary dramatically in terms of length. The long comment is focused on substance, is clear regarding the problem's location, explains why the content is incorrect, and provides additional related ideas. The short comment is focused on surface-level writing, is vague about the nature of the problem, is unclear about the problem's location, and makes no suggestions for making changes. A long comment does not guarantee that all useful elements of a comment are present, but useful elements are statistically much more likely to be present in such long comments (Wu & Schunn, 2020).

Reviewer	Your Comment(s)	Reviewer #2	Reviewer #3	Reviewer #4
Concise writing	I think the writing is great, just finish it *Backevaluation(2)*: Obviously.	Well, considering you are not finished, it was pretty good. I have a few suggestions. "we have to look at the relationship between the brain and memories, as well as the relationship with strategies as the ones mentioned above. In our class, we discussed what exactly remembering and forgetting was. Is remembering just finding the copy of a memory or just rebuilding it? Is forgetting just losing a memory for good or one that hasnâ??t been found yet? We theorize a lot about things like this and thus, we study how we can better ourselves in learning how to remember things through research. An important start for this is organization, context and the overall importance of structure. There was a study done on the impact of structure, one which we discussed in one of our lectures, that detailed how recollection was affected by organized or random lists." I would change the part above by: adding a transition between the two paragraphs "we have a lot of different theories regarding this topic so we can study how we can better ourselves in learning how to remember. The bases of this topic is organization, context, and the overall importance of the structure. Structure can be defined as____. A study on this topic found that recollection was affected by organized or random lists. *Backevaluation(5)*: This was very helpful! I know sometimes my sentence and paragraph structures can be out of sorts, but sometimes I don't know where to fix it. But this specific example and suggestion will help me a lot!	You do a good job keeping the essay concise. It never really seemed needlessly wordy to me. *Backevaluation(4)*: Straight forward	Overall, the author uses an appropriate amount of detail. The introduction is heavily detailed, however, this may be a stylistic choice as the author is trying to form a narrative. *Backevaluation(5)*: Introduction was heavy in detail, so I might try to make it more concise or make it shorter.

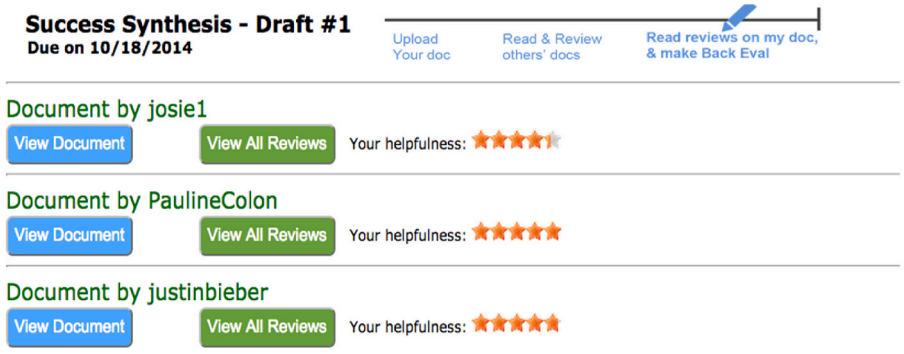

Fig. 3. Top: The Peerceptiv interface for seeing helpfulness ratings (within parentheses) and supporting explanations given to each comment. Bottom: The interface showing the summary of helpfulness ratings for each completed review (averaged across reviewing dimensions).

We defined long comments as those having at least 50 words. This threshold allows for a comment to have a problem description, problem location, problem explanation, and possible solution. Further, it clearly eliminates all highly superficial comments (e.g., vague feedback or praise-only feedback). It captures an obtainable minimum length in those students regularly provided at least a third of the time. Finally, an even higher threshold might be challenging to reach for some lower-level language dimensions. To inform the selection of this specific threshold and to validate the overall measurement approach, variations using different thresholds were created, and then the correlation with helpfulness ratings was examined. The 50 words had the highest correlation (see Table B1 in Appendix B). Further, pilot analyses suggested that changing the threshold higher or lower did not substantially change the main regression results in this study.

Two measures of reviewing participation were developed based upon this long comment threshold, both representing the pattern of long comments across the reviews provided on a given assignment. First, we calculated *#Long provided*$_J$, the total number of long comments a reviewer gave across reviewing dimensions and documents to be reviewed on assignment *J*. This measure is an assessment of how much in-depth reviewing a reviewer was providing. Note the convention of using # as a symbol for "number".

Second, we calculated *%Long provided*$_J$, the percentage of all provided comments by a reviewer for assignment *J* that were long. The number of long comments can be influenced by either a tendency to provide in-depth comments or a decision to give more comments. The latter might be heavily influenced by the quality of the documents that were reviewed.

To examine the reliability of these measures, scores on each measure were re-calculated at the level of an individual review, and then Cronbach's alpha was calculated across the minimum number of reviews that each student provided (i.e., four reviews). Both measures showed high reliability (see Table C1 in Appendix C).

4.1.2. *Experienced inequality*

There were two measures that directly reflected two forms of inequality students experience in contrast between what they provided as reviewers and what they received as authors. First, capturing kind of qualitative inequality, there was *Rating inequality*$_J$, defined as the difference between the mean helpfulness ratings on comments provided and the mean helpfulness ratings on comments received on assignment *J*. Thus, positive numbers reflect the provided comments were more

Table 2
The specific measures for each construct and their definitions.

Constructs	Measures	Definition
Reviewing participation	*#Long provided$_J$*	The number of comments provided by student *i* on the J^{th} assignment that were long (#words>50).
	%Long provided$_J$	The percent of comments provided by the student *i* on the J^{th} assignment that were long (#words>50).
Experienced inequality	*Rating inequality$_J$*	For student *i* on the J^{th} assignment: (mean helpfulness of comments given minus mean helpfulness of comments received)/standard deviation of rating inequalities on that assignment
	#Inequality$_J$	For student *i* on the J^{th} assignment: (number of long comments given minus number of long comments received)/standard deviation of number inequity on that assignment
Task performance	*Low$_J$*	1 if the document score on the J^{th} assignment is lower than the median score, and 0 otherwise (higher or not submitted)
Within-course experience	*Round*	The reviewing assignment number ($J = 1, 2, 3, 4, \ldots$)
Course type	*Discipline*	Indicator variable: set to 1 for psychology, 0 for biology
	Level	Indicator variable: set to 1 for the advanced course, 0 for the introductory courses

Table 3
Means and standard deviations in each course for each predictor and outcome variable. Note that raw inequality values are presented in the table, but standardized inequality values were used in analyses (i.e., necessarily had mean = 0 and SD = 1).

Variable	Introductory Psychology		Introductory Biology		Advanced Biology	
	Mean	SD	Mean	SD	Mean	SD
Predictor						
#Long provided$_{J-1}$	7.48	6.94	2.58	3.14	7.48	7.42
%Long provided$_{J-1}$	34%	28%	32%	31%	35%	27%
Raw Rating inequality$_{J-1}$	0.00	0.65	0.11	2.15	0.08	0.83
Raw #Inequality$_{J-1}$	−0.37	11.62	0.00	7.24	−0.12	12.47
Low$_J$	0.40		0.41		0.44	
Round	2.79	1.33	2.99	1.41	2.50	1.12
Outcome						
#Long provided$_J$	6.82	6.61	2.72	3.32	8.41	7.18
%Long provided$_J$	31%	28%	32%	32%	36%	26%

helpful, and negative numbers reflect the received comments were more helpful. Note that cases in which authors did not provide helpfulness ratings are treated as missing values and dropped from mean helpfulness calculation. Second, capturing a kind of quantitative inequality, there was *#Inequality$_J$*, defined as the difference between the total number of comments provided and the total number of comments received on assignment *J*. Correspondingly, positive numbers reflect more comments given, and negative numbers reflect more comments received. Note these two measures directly capture experienced inequality, and there is no concern about measurement noise and reliability for such measures (e.g., similar to direct measures of income or income inequality). In addition, these variables were standardized (mean centering and setting variation to 1) for the regression models to allow for testing of interactions.

4.1.3. Task performance

Document quality was based upon the multi-peer ratings. The mean inter-rater reliability on a given rating rubric in these three courses was 0.57. There were between 6 and 14 ratings per assignment. Thus, interrater reliabilities for the overall document scores were generally high (i.e., always above 0.8). In general, with clear rubrics and multiple peer ratings, a prior meta-analysis indicates that the validity of these ratings was likely to be high (Li et al., 2016). However, the exact validity of these ratings across such a wide variation of assignments across courses was not possible to assess formally. Therefore, document quality was formally modeled in terms of a binary measure, *Low score$_J$*, which was set to 1 if the score on assignment *J* was below the class median for that assignment and 0 otherwise (i.e., received a high score). This transformation to a binary value also handled the few cases in which no document was submitted as well as addressing outliers and the skew found in score distributions for most assignments.

4.1.4. Within course experience

To capture quantitative changes in reviewing performance changes across the course (e.g., become less sensitive to inequality later in the course), time was formally modeled in terms of the amount of experience with reviewing. In particular, *Round* was defined as the assignment number in the course (and it is equivalent to *J*).

4.1.5. Course type

A *Discipline* indicator variable was created to differentiate the different course disciplines: 1 for psychology and 0 for biology. A *Level* indicator variable was created to distinguish the introductory courses (0) from the advanced course (1).

4.2. Analysis

The general analytic approach uses multiple regression to test the statistical significance of predictors that correspond with each of the hypotheses. In particular, we use time-series multiple regressions in which reviewing participation in one assignment is predicted by experienced inequality and task performance in the prior assignment, controlling for reviewing participation in the prior assignment. This approach of predicting the outcome variable while controlling for the prior state is preferred over predicting change scores (current minus prior) because change scores are subject to regression-to-the-mean statistical artifacts in multiple regression analyses (Barnett et al., 2005). Thus, the specific analytic method was to use multiple regression, with the number of long comments provided (*#Long provided$_J$*) or the percent of long comments provided (*%Long provided$_J$*) as the dependent variables, *#Long provided$_{J-1}$* or *%Long provided$_{J-1}$* as the respective baseline controls, *Number inequality$_{J-1}$*, *Rating inequality$_{J-1}$*, as core predictors, and *Low$_J$*, *Round*, *Discipline*, and *Level* as additional control variables.

The nature and observed distributions of the two outcome measures of reviewing participation, *#Long provided$_J$* and *%Long provided$_J$*, were important to consider in specifying the details of the multiple regression approach. Both variables' distributions were far from normal (see Figure D1 in Appendix D). *#Long provided$_J$* is a count variable, which tends to have a strong positive skew (i.e., a long tail on the right). In such situations, Poisson regression or Negative Binomial regression is preferred over linear regression based upon a normal distribution. Since the variance for the number of provided comments was generally much larger than its mean in each course (see Table 3), it was likely that Negative Binomial regression would be the better choice (Grogger & Carson, 1991; Gardner et al., 1995). As expected, Negative Binomial regression produced better model fit statistics than did Poisson Regression (Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood) and thus is the approach reported here.

Another problem with some count variables is that they will also have many zeros when there is a theoretically meaningful, separate process for producing zeroes. In providing comments, some reviewers can choose not to provide any comments to peers in a given round, which could represent a meaningful reaction to inequality in prior reviewing experiences (as well as other stressors like competing course deadlines or sickness). In such a situation, Zero-Inflated regression should be used, in which two sets of coefficients are produced: a logistic regression predicts zeroes' occurrence in the outcome variable (i.e., no long comments provided); and another regression (in this case, Negative

Binomial regression) predicts the quantity when it is not zero.

%Long provided$_J$ is a percent outcome, which often can have other kinds of statistical issues, particularly related to overdistribution (i.e., a very wide and almost flat distribution) and a peak in the number of values at the endpoints of the scale due to truncation (i.e., at 0 or 1). In the case of *%Long provided$_J$*, the distribution was very flat, and there was a peak at 0 (see Figure D1). Ignoring these normality violations produces artificially depressed correlations (Hammer & Landau, 1981) or noisy estimates (Maddala, 1983). Tobit regression (Tobin, 1958) can address these issues and produce more consistency, reliability, and less bias (Maddala, 1983; Leigh et al., 1985). Therefore, Tobit regression was selected for the regressions predicting *%Long provided$_J$*.

Initial explorations found roughly linear relationships between predictor variables and outcome variables, with the exception of task performance. Therefore, the core predictors are left as linear predictors, and task performance (*Low$_J$*) was modeled as a binary (Low or not). Data from the different courses are combined into one regression model. Control variables adjust for overall differences by course, and additional models were conducted to formally test the course variables' interactions with the core predictor variables. Further, interactions with *Round* and task performance (*Low$_J$*) were also formally tested in interaction models. The core predictors (rating and number inequality) were standardized to avoid artificial effects in the models testing interaction terms (Baguley, 2009).

For all models (Zero-Inflation Negative Binomial (ZINB), and Tobit regression models), Pseudo-R^2 is reported to show model quality. Outliers in each continuous predictor were replaced with the closest non-outlier value; such outliers occurred for at most 0.4% of values for any given variable and less than 0.1% of values for most variables.

5. Results

The top of Table 4 shows the Pearson correlations among the predictors. Except for the two baseline variables, which are never included in the same regression, none of the predictors were strongly related to each other, suggesting major multicollinearity problems would be unlikely. Inspection of Variance Inflation Factors (VIF) in the regression models confirmed there were no multicollinearity problems (i.e., all VIFs <2).

The bottom of Table 4 shows a strong relationship of the baseline variables to the outcome variables, but not so strong as to prevent predicting important variance in change over time. Further, prior rating inequality was a consistent negative correlate of both length outcomes and both length baselines, so a regression analysis is needed to tease out predicting future change from concurrent relationships. Number inequality was only related to the number but not percent long variables at baseline or outcome.

Current low task performance correlated significantly with the outcome measures and the core predictors, justifying its inclusion as a control variable. *Round*, *Discipline*, and *Level* were not important in this way (e.g., not correlated with the core predictors) but were kept as control variables because they turned out to be related to the outcome when other controls were included and because interactions with *Round* and *Level* were key hypothesis tests.

5.1. Predicting number long

Table 5 presents the ZINB regression results of the baseline (only baseline and control variables) and inequality (adds both inequality predictors) models for the dependent variable *#Long provided$_J$*. Note that in this table and other ZINB regression tables, the sign of the coefficients for the zero-inflation component is reversed (i.e., predicting whether the student provided at least one comment rather than predicting when they provided zero comments) to avoid having to interpret double-negatives.

The baseline variable is a significant predictor for both the number of comments and providing at least one comment, highlighting the importance of controlling for past behavior. Most importantly, in partial support of Hypothesis 1, both rating inequality and number inequality predictors were significantly related to the number of long comments provided in the next assignment. That is, students provided fewer comments in the reviewing assignment when they experienced rating inequality or number inequality in the prior assignment. However, there was no relationship of rating or number inequality with whether any comments were provided in the next assignment. All four control variables were also statistically significant and therefore important factors to include.

The number inequality and rating inequality variables were each recoded into five categorical levels (Very low, Low, Medium, High, Very high) in order to visualize the associations. Roughly, the two high levels reflect cases in which students provided more (or more helpful) comments than they received, whereas the two low levels reflect cases in which they received more (or more helpful) than they provided. Then, estimated marginal means for the number of long comments provided were calculated for each categorical level from a regression controlled for baseline levels, the other inequality variable, and the various control variables. Fig. 4 (left) shows a graph of these means. Both inequality variables show a small but consistent gradual negative association of inequality with the number of long comments provided in the next assignment. Notably, the negative associations with giving too much appeared to be larger than the positive associations with receiving too much. Taking Medium as the neutral point (where amount or quality provided roughly matched the amount or quality received), Very High experienced inequality was associated with giving roughly one less long comment in the next assignment, whereas Very Low experienced inequality was associated with barely any difference in the next assignment.

Estimated marginal means were also calculated for the percentage of students submitting a long comment using the same five levels of the rating and number inequality variables. As shown in Fig. 4 (right), there were small negative trends for inequality in whether the student submitted any long comments at all, but with higher variability at the Very Low and Very High endpoints. This larger amount of noise at the

Table 4
Pearson intercorrelations among the baseline, core, and control predictors (top) and with the outcome variables (bottom).

	#Long provided$_{J-1}$	%Long provided$_{J-1}$	Rating inequality$_{J-1}$	# Inequality$_{J-1}$	Low$_J$	Round	Disc.	Level
#Long provided$_{J-1}$								
%Long provided$_{J-1}$	.73***							
Rating inequality$_{J-1}$	-.25***	-.27***						
#Inequality$_{J-1}$	.34***	-.03	-.15***					
Low$_J$	-.17***	-.15***	.08***	-.10***				
Round	.03	.03*	-.03	-.01	.03			
Discipline	.11***	.01	.01	.002	-.05**	.06**		
Level	.08***	.02	-.01	-.002	.05**	-.13***	-.74***	
#Long provided$_J$	.60***	.49***	-.19***	.17***	-.15***	-.13***	.01	.18***
%Long provided$_J$	.50***	.59***	-.19***	.003	-.14***	-.09*	-.06***	.08***

Notes. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 5
Estimated coefficients (for both count and non-zero outcomes) from the ZNBR baseline and inequality predictor models for #Long provided$_J$ as the outcome variable, along with model fit statistics.

Predicting *#Long provided*$_J$	Model 1 (Baseline)		Model 2 (Inequality)		Model 3 (Level X)		Model 4 (Round X)		Model 5 (Low X)	
	Count	Not Zero	Count	Not Zero	Count	Not Zero	Count	Not Zero	Count	Not Zero
Baseline										
#Long provided$_{J-1}$	0.06***	0.42***	0.06***	0.43***	0.06***	0.42***	0.06***	0.41***	0.06***	0.41***
Core predictors										
Rating inequality$_{J-1}$			−0.05**	−0.07	−0.05**	−0.08	−0.04**	−0.07	−0.05***	−0.05
#Inequality $_{J-1}$			−0.04**	−0.02	−0.04**	0.06	−0.04***	−0.04	−0.04**	−0.05
Level x Rating inequality$_{J-1}$					0.02	−0.17				
Level x #Inequality$_{J-1}$					−0.08**	0.52*				
Round x Rating inequality$_{J-1}$							0.02	−0.11*		
Round x #Inequality$_{J-1}$							0.003	−0.09		
Low$_J$ *x Rating inequality*$_{J-1}$									−0.03	−0.08
Low$_J$ *x #Inequality*$_{J-1}$									0.03	−0.25
Control variables										
Round	−0.10***	−0.15**	−0.10***	−0.14*	−0.10***	−0.13*	−0.10***	−0.09	−0.10***	−0.14*
Low$_J$	−0.08**	−0.51***	−0.08**	−0.49**	−0.08**	−0.50***	−0.08**	−0.49***	−0.08**	−0.41*
Discipline	0.37***	−0.03	0.38***	−0.00	0.37***	0.002	0.39***	0.00	0.38***	0.02
Level	0.59***	1.06***	0.58***	1.16***	0.58***	1.43***	0.58***	1.18***	0.58***	1.18***
N	3029	2542	2869	2436	2869	2436	2869	2436	2869	2436
Pseudo R^2	0.07	0.22	0.07	0.23	0.07	0.22	0.07	0.23	0.07	0.23

Notes. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

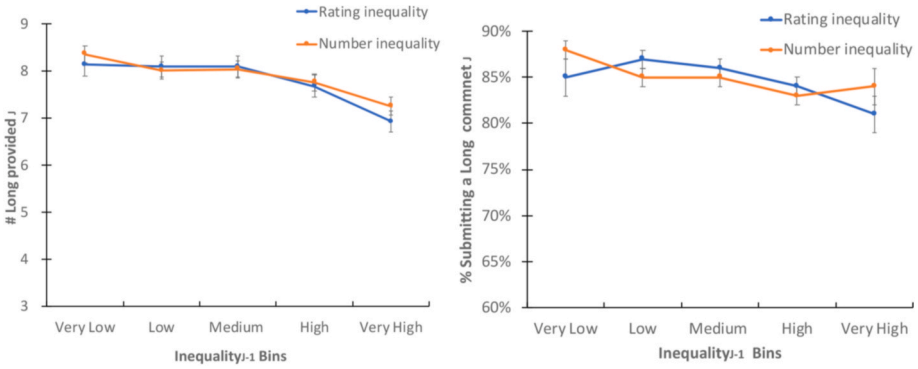

Fig. 4. The estimated marginal means (with SE bars) for *#Long provided*$_J$ (left) and % of students submitting at least one long comment on the *J*th assignment (right) as a function *Rating inequality* and *#Inequality* in the prior assignment, after controlling for baseline and control variables.

endpoints may explain why the associations were not significant overall for this zero-inflation part of the model.

Table 5 also presents the results of the additional regressions that tested the hypothesized interactions of rating and number inequality with course level (Model 3), with round (Model 4), and Low task performance (Model 5). Note that interactions with course discipline were also explored, but no significant interaction was found. There was no significant interaction of inequality with Low Task performance (Hypothesis 2).

Regarding interactions with *Round* (Hypothesis 3), there was a significant interaction in the case of zero-inflation for rating inequality. To visualize this association, we again recoded rating inequality (this time into two categories to reduce noise because zeroes were relatively rare) and also implemented a median split on *Round* (first half of assignments vs. second half of assignments in each course). Then, we graphed the estimated means from the interaction while including baseline and control variables (see Fig. 5). In early assignments, there was no association of rating inequality with the likelihood of submitting at least one long comment. However, in the later assignments (consistent with Hypothesis 3), students were approximately 5% less likely to submit any long comments after experiencing high rating inequality.

There was a statistically significant interaction of number inequality with course level (Hypothesis 4). To visualize that interaction, a similar approach with the data was used by recoding inequality levels, but separately for each course. Because there was much less data within just one course and the overall association appeared linear, only three larger categories were used. Consistent with the regression results, while rating inequality had similarly strong associations with provided comments in all three courses (see Fig. 6 right), there appeared to be a more muted and non-significant association with number inequality in the *Advanced Biology* course (see Fig. 6 left), counter to Hypothesis 4.

5.1.1. *Predicting percent long*

Table 6 presents all five regression models for the Tobit regression models of *%Long provided*$_J$. Here, the results partially supported Hypothesis 1: experiencing higher Rating inequality in the prior assignment predicted less reviewing participation (i.e., a lower percentage of comments provided that were long) in the current assignment. This relationship held in the main effects model (model 2) and in each of the interaction models (models 3 through 5). However, number inequality was not a significant predictor for this outcome variable in any of these models.

Turning to the moderators, there was a significant positive interaction of number inequality with low task performance on the current assignment. This association was graphed using the median splits for both number inequality and task performance. As shown in Fig. 7, the same trend was shown for high vs. low prior performance, but the negative association with greater number inequality was larger in the low prior performance group, consistent with Hypothesis 2. There was no significant interaction with Round, contradicting Hypothesis 3.

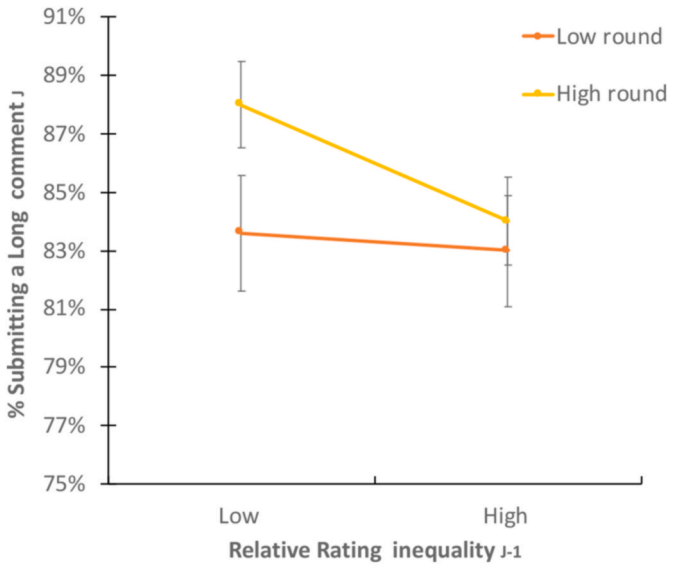

Fig. 5. Separately for early vs. later assignments (round), the estimated percentage of students submitting at least one long comment on the J^{th} assignment (with SE bars) as a function of experiencing relatively low vs. high *Rating inequality* in the prior assignment, after controlling for baseline and control variables.

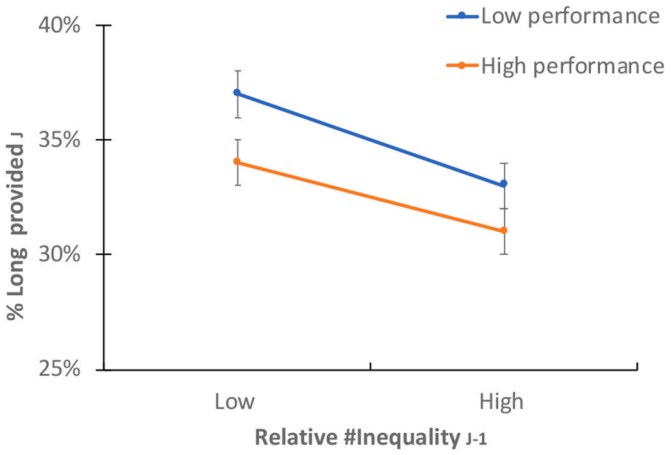

Fig. 7. Separately for students with high or low prior assignment performance, the mean estimated percentage of comments that were long on the J^{th} assignment (with SE bars) as a function of experiencing relatively low vs. high number inequality in the prior assignment, after controlling for baseline and control variables.

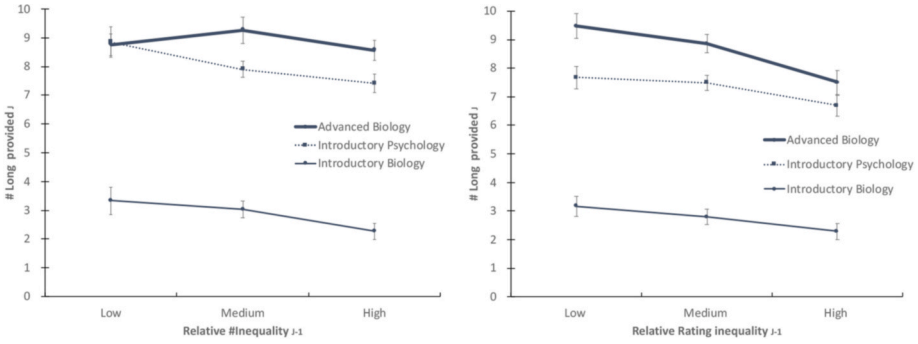

Fig. 6. The estimated marginal means (with SE bars) for *#Long provided*$_J$ (controlling for *#Long provided*$_{J-1}$) as a function relative levels of number inequality (left) and rating inequality (right) in each of the three courses, after controlling for baseline and control variables.

Table 6
Estimated coefficients from the Tobit baseline, inequality, and interaction predictor models for *%Long provided*$_J$ as the outcome variable, along with model fit statistics.

Predicting *%Long provided*$_J$	Model 1 (Baseline)	Model 2 (Inequality)	Model 3 (Level X)	Model 4 (Round X)	Model 5 (Low X)
Baseline					
%Long provided$_{J-1}$	0.66***	0.64***	0.64	0.64***	0.64***
Core predictors					
Rating inequality$_{J-1}$		−0.01**	−0.01**	−0.01**	−0.01**
#Inequality$_{J-1}$		0.00	0.00	0.00	0.004
Level x Rating inequality$_{J-1}$			0.02		
Level x #Inequality$_{J-1}$			0.05***		
Round x Rating inequality$_{J-1}$				−0.001	
Round x #Inequality$_{J-1}$				0.00	
Low$_J$ *x Rating inequality*$_{J-1}$					−0.003
Low$_J$ *x #Inequality*$_{J-1}$					0.02*
Control variables					
Low$_J$	−0.04***	−0.04***	−0.04***	−0.04***	−0.04***
Round	−0.03**	−0.03***	−0.03***	−0.03***	−0.03***
Discipline	0.00	0.00	0.00	0.00	0.00
Level	0.05***	0.06***	0.06***	0.06***	0.06***
N	2930	2815	2815	2815	2815
Pseudo R^2	0.52	0.54	0.54	0.54	0.54

Notes. *** = p < .001, ** = p < .01, * = p < .05.

Course level again shows significant interaction with number inequality, but this time with a larger association in the advanced course, which was consistent with Hypothesis 4.

6. General discussion

This study's main aim was to test whether (and when) students' inequality experiences in peer review were associated with the depth of their subsequent participation as predicted by a game theory analysis, focusing on the role of number and rating inequality. We explored whether the willingness to make knowledge contributions changed with different degrees of experienced interactive inequality. In particular, we tested four hypotheses about the overall relationship of inequality with participation and three kinds of potential moderators that could increase the association of inequality with participation. Table 7 summarizes the findings for each of the four hypotheses.

Overall, Hypothesis 1 received consistent support. Both rating inequality (a qualitative inequality) and number inequality (a quantitative inequality) were associated with providing fewer long comments in the next assignment. The relationship was generally more consistent in terms of the number of long comments provided rather than whether any long comments were provided or the percent of long comments provided. In any case, the current study is consistent with broader research on game theory and similar theories applied to social information exchange, especially with respect to interaction inequality driving community participation (Civai et al., 2010; Yang & Ott, 2016; Fieseler et al., 2019).

In addition, the predictiveness of experienced inequality was moderated by all three hypothesized factors, also similar to the broader research literature, which finds moderation of inequality effects by contextual factors (Hu et al., 2016; Khandeparkar et al., 2020). However, in some cases, the moderation was only quantitative and weak. When students were performing at lower levels in the assigned task, they appeared more sensitive to rating inequality than when they performed at a higher level (Hypothesis 2). The quantitative nature of this moderation may explain why it was not found to be supported for the number of comments as the outcome and only partially supported when the percentage of comments that were long was the outcome.

Why would task performance moderate inequality? One plausible explanation is that students with lower performance tend to have less confidence in their knowledge, particularly related to the current assignment. This lower confidence would lead them to produce fewer comments to their peers, especially fewer long comments that contain suggestions for improvements. If they also experience inequality, they may come to believe that long comments are not normatively required.

In the other two cases, the moderation was substantial and qualitative such that the predictiveness of experienced inequality for future participation only occurred in one situation. For example, inequality appeared to predict only in later assignments and not in earlier assignments (Hypothesis 3). Similarly, the course level appeared to substantially moderate the predictiveness of inequality (Hypothesis 4), although in complex ways. These opposite patterns in course-level interactions across outcome variables might explain each other. That is, it may be that in all courses, students react negatively to number inequality, but in different ways. Students may have changed their reviewing style to be less commonly long in the advanced course, whereas students reacted by provided fewer comments in the introductory courses (short and long).

Table 7
Findings summary for each hypothesis for number or percent outcome variables (along with whether key predictors for each hypothesis of rating inequality, number inequality, or both were statistically significant).

Hypothesis	Outcome	
	#Long provided	%Long provided
H1: Experiencing reviewing inequality (either number or rating) in the prior peer review assignment will be associated with lower future reviewing participation.	Supported (Both)	Supported (Rating)
H2: The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified when students do poorly in the course.	Not Supported	Supported (Number)
H3: The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified with later peer review assignments.	Supported (Rating)	Not Supported
H4: The negative associations of experiencing reviewing inequality with future reviewing participation will be magnified in more advanced courses.	Not Supported (Opposite for Number)	Supported (Number)

6.1. Theoretical implications

This study contributes several theoretical insights. First, the study empirically establishes the relevance of game theory to computer-based peer review beyond the preliminary work on how peers grade each other and whether they accept the feedback they receive. Early research on peer review was strongly focused on the accuracy of peer assessment, which was generally found to be good in meta-analyses (Li et al., 2016). Later research focused on learning benefits, which have been consistently confirmed across four recent meta-analyses (Huisman et al., 2019; Double et al., 2019; Li et al., 2020; Zheng et al., 2020). The current study of equality effects complements these prior lines of research because both the accuracy and learning benefits of peer review depend upon strong participation in peer review. The current work shows some sensitivity to imbalances in participation in peer feedback, which is a natural extension of the large body of Game Theory research showing people pervasively react to competitors' behaviors in repeated decisions.

Second, the study was the first to bring together qualitative and quantitative aspects of inequality. By carefully controlling for each other within the regression models, the study suggests both dimensions are indeed important to inequality. Further, since one effect is based in objective behaviors and the other is based in subjective perceptions, the study essentially provides converging evidence for the importance of inequality in computerized peer review.

Third, this study suggests some asymmetry between the positive and negative sides of inequality in its relationship to future behaviors. In particular, the study suggests that the negative effects of inequality (i.e., providing more than receiving) outweigh the positive effects of reverse inequality (i.e., receiving more than providing). This finding is consistent with other research on received injustices weighing particularly heavily (Heerink et al., 2001; Oishi et al., 2011).

Finally, this study examined theoretically predicted moderators of inequality that were not previously tested. For-whom and under-what-circumstances are important questions for judging the broader value and generality of theories as well as providing further support for their hypothesized underlying mechanisms. For example, the moderating role of student performance could be conceptualized as speaking to the underlying perception of equality: students with greater skills neither need more help nor need to work as hard in order to provide comments, and thus it may be perceived as equal when they provide more long comments.

6.2. Practical implications

We first begin with a discussion of effect sizes. It is important to note that the effects of inequality in just one round were relatively small in size. As simple models are applied to more micro-level processes (e.g., reviewing behaviors within an assignment), more local factors (e.g.,

competing time constraints, quality of reviewed objects) play a larger role, and the variance explained by the model focused on developmental patterns is generally predicting small amounts of change. However, across multiple rounds of reviewing, such effects can accumulate, particularly when there are positive feedback loops. For example, a very conscientious reviewer can gradually come to produce very brief or superficial comments. Indeed, a whole class can gradually come to produce few and superficial comments as the cases of inequality build over time. Thus, the small micro-patterns observed in this study can be argued to have several practical implications.

First, as noted above, it appears there is a bigger negative association with high inequality than a positive association with low inequality. This suggests instructors should intervene early to reduce the likelihood that students experience quantitative or qualitative inequality because the combined effects of natural variation in the amount/quality of feedback appear to be negative. For example, clear guidance and training (Rae et al., 2014; Rodríguez-Gómez et al., 2016) on how many comments should be provided and what kinds of content should be found in each comment should reduce the levels of experienced inequality. Simple automated systems could also be created that warn feedback providers (or instructors) that some reviews contain too little feedback (Ramachandran et al., 2017). Luckily, the effects of inequality appear to be larger in later assignments such that the instructor can develop better guidance for reviewers during the course.

Another observation that is also potentially relevant to practice is that both qualitative inequality and quantitative inequality appear to matter. Therefore, educators should not only emphasize the number of feedback comments provided. It is tempting to focus on the number of comments because it is easy to provide such guidance to students or set up a system that automatically checks the number of comments provided (e.g., within Turnitin, Peerceptiv). But requirements focused only on the amount of feedback may lead students to provide meaningless comments simply to meet course requirements, which affect experienced inequality and limit what students will learn from providing and receiving feedback (Wu & Schunn, 2020).

Finally, the relationship of inequality to future providing behaviors appears to be larger for struggling students, which suggests that more attention should be paid to those cases. For example, instructors could ensure that weaker students' contributions are assigned to reviewers who tend to provide better comments and provide additional comments when such students appear to have received too little feedback so far. Random assignment of reviewers to submissions is common (Babik et al., 2016), but this may too often lead students with low task performance to spiral downwards into low levels of participation. Such an effect is unfortunate given the pedagogical advantage of providing feedback to peers (Li et al., 2020; Wu & Schunn, 2020).

7. Limitations and future research

It is important to acknowledge several limitations in the current study, which in turn provide directions for future research. First, this study only focused on rating and number inequality. These were found to be important predictors, but other kinds of experienced inequality might matter as well. For example, in situations where feedback is distributed continuously rather than all at once, feedback timeliness may be unequal. Alternatively, the constructiveness (e.g., containing possible solutions) or "on-taskness" (addressing the core elements of the assignment) or negativity (especially strong negative comments) of feedback may also be important to perceptions of equality (Ramachandran et al., 2017). Future work could use a qualitative approach to uncover the commonly mentioned dimensions of inequality in different peer review settings.

Second, it is important to note that the tests of Hypothesis 4 are particularly preliminary in the current study. There were a number of differences between the courses that might also account for the observed variation. For example, it is unclear whether these differences are due to classroom norms or the amount of variation in student ability, or the nature of the assignments. It may also be relevant that the absolute number of long comments was much higher in the Advanced course, as was variability in the number of long comments provided (see Table 3). A broader range of courses must be studied to more rigorously test the effects of course type. At this point, the exploration of the relationship of inequality with review participation across three courses provides support for the pervasiveness of inequality effects and some suggestions that the size of the relationship will vary across courses.

Third, the study only focused on computer-based peer review in classrooms. Theoretically, the same factors should also matter for other forms of online peer review, such as conference, journal, or grant reviewing, as well as other kinds of electronic information exchanges such as in *reddit* or *Yahoo Answers*. The current study provides a methodological model for studying inequality effects in naturalistic ways in those other forms of peer review or information exchange, especially with respect to quantitative inequality.

Fourth, training or system factors may moderate the effects of interaction inequality. Replicating the study across different peer review systems will be necessary, particularly ones that substantially vary their support for quality reviewing or contexts that use much more or much less training on peer feedback. It will also help us understand when interactional inequality tends to be common and when it tends to influence reviewer participation.

Fifth, while this study combined objective behavior and subjective survey measures to derive estimates of experienced inequality, it did not directly include subjective measures of perceived (in)equality, as is often the case in research based on the game theory paradigm. From a pragmatic perspective, it may not matter whether students were explicitly aware of the inequality; experience can implicitly shape action as well (Reder & Schunn, 1996; Christiansen, 2019). However, future studies that specifically measure perceptions of equality in reviewing would be helpful to more precisely characterize the mechanisms underlying the relations observed in this study.

Finally, this study was inherently a correlational study, and thus the causal claims are somewhat limited. The longitudinal nature of the analyses (i.e., how prior experience relates to future performance) rules out reverse causal interpretations, and the multiple regression techniques controlled for several potentials confound. However, a study that directly manipulated the inequality of feedback (e.g., by withholding some of the feedback to a random subset of authors) would be a good complement to the naturalistic study reported here. Naturalistic studies that capture human behaviors in computerized contexts provide important external validation of phenomena (Lerche & Kiel, 2018; Li et al., 2013), particularly in ways that are non-intrusive. Manipulations and extensive research-based measurement have the potential of modifying behaviors in artificial ways.

Credit author statement

Zheng Zong: Conception proposal, Data acquisition, Analysis, Visualization, Original draft writing; Christian D. Schunn: Conception proposal, Work design, Data source, Interpretation of data; Yanqing Wang: Proofreading, Reviewing, Correspondence.

Funding

This work was partially supported by the National Natural Science Foundation of China [71573065].

Appendix A

Table A1
Means and standard deviations in Introductory Psychology course for each predictor and outcome variable.

Variable	Assignment 2		Assignment 3		Assignment 4		Assignment 5		Assignment 6	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Predictor										
#Long provided$_{J-1}$	8.39	7.43	8.00	6.91	7.24	6.56	6.80	6.54	6.41	7.17
%Long provided$_{J-1}$	38%	29%	39%	30%	33%	27%	29%	26%	28%	27%
Raw Rating inequality$_{J-1}$	0.03	0.65	0.02	0.68	−0.03	0.59	−0.02	0.63	0.00	0.73
Raw #Inequality$_{J-1}$	−0.24	10.94	−0.21	10.56	−0.32	11.55	−0.76	13.41	−0.28	11.38
Low$_J$	0.51		0.53		0.54		0.58		0.35	
Outcome										
#Long provided$_J$	8.00	6.91	7.24	6.56	6.80	6.54	6.41	7.17	4.98	6.31
%Long provided$_J$	39%	30%	33%	27%	29%	26%	28%	27%	21%	25%

Table A2
Means and standard deviations in Introductory Biology course for each predictor and outcome variable.

Variable	Assignment 2		Assignment 3		Assignment 4		Assignment 5		Assignment 6	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Predictor										
#Long provided$_{J-1}$	2.21	3.07	2.54	3.59	2.73	3.04	2.89	3.09	2.52	2.88
%Long provided$_{J-1}$	23%	28%	31%	32%	34%	31%	37%	34%	35%	32%
Raw Rating inequality$_{J-1}$	−0.10	1.69	0.12	2.20	0.28	2.19	0.12	2.11	0.10	2.53
Raw #Inequality$_{J-1}$	0.00	8.11	0.00	7.42	0.00	6.70	0.00	6.98	0.00	7.06
Low$_J$	0.40		0.42		0.42		0.40		0.42	
Outcome										
#Long provided$_J$	2.54	3.59	2.73	3.04	2.89	3.09	2.52	2.88	2.83	3.93
%Long provided$_J$	31%	32%	34%	31%	37%	34%	35%	32%	32%	33%

Table A3
Means and standard deviations in Advanced Biology course for each predictor and outcome variable.

Variable	Assignment 2		Assignment 3		Assignment 4		Assignment 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Predictor								
#Long provided$_{J-1}$	2.54	3.31	8.15	6.38	10.98	10.17	8.28	5.23
%Long provided$_{J-1}$	23%	28%	34%	24%	33%	24%	49%	27%
Raw Rating inequality$_{J-1}$	0.17	0.95	0.11	0.80	0.07	0.73	0.01	0.82
Raw #Inequality$_{J-1}$	−0.15	7.43	−0.13	8.71	−0.10	21.12	−0.11	6.72
Low$_J$	0.41		0.44		0.45		0.46	
Outcome								
#Long provided$_J$	8.15	6.38	10.98	10.17	8.28	5.23	6.47	4.79
%Long provided$_J$	34%	24%	33%	24%	49%	27%	35%	25%

Appendix B

Table B1
Pearson intercorrelations of mean helpfulness ratings with the proportion of comments that are long based upon different length thresholds for defining a long comment.

Threshold Defining Long Comments	Correlation with Helpfulness rating
10 words	0.08
30 words	0.18
50 words	0.20
100 words	0.12

Appendix C

Table C1
Item reliability statistics and overall scale Cronbach alpha values for each outcome variable based upon contributions of the variable defined separately on each of the first four completed reviews on each assignment.

	N	item-test correlation	item-rest correlation	average interitem covariance	alpha
#Long measure					
*1*st	2886	.884	.775	1.33	
*2*nd	2926	.883	.775	1.28	
*3*rd	2802	.879	.781	1.34	
*4*th	2734	.837	.711	1.56	
SCALE					.894
%Long measure					
*1*st	2487	.864	.721	1.05	
*2*nd	2429	.857	.722	1.03	
*3*rd	2312	.847	.715	1.04	
*4*th	2245	.837	.701	1.03	
SCALE					.859

Appendix D

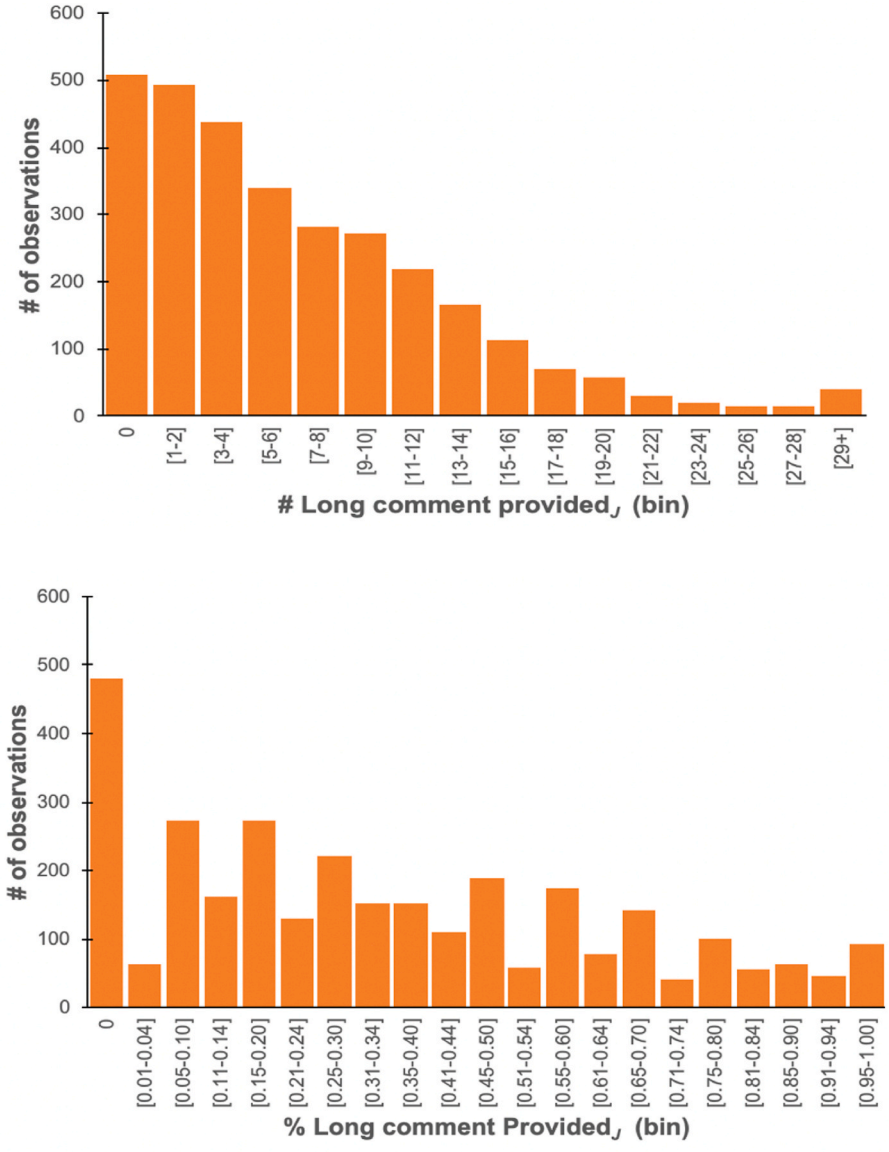

Fig. D1. Frequency histograms for the number of comments provided and percent of comments provided that are long on the J^{th} assignment.

References

Anbarci, N., & Feltovich, N. (2013). How sensitive are bargaining outcomes to changes in disagreement payoffs? *Experimental Economics, 16*(4), 560–596.

Applebee, A., & Langer, J. (2011). *The national study of writing instruction: Methods and procedures*. Albany, NY: Center on English Learning & Achievement.

Babik, D., Gehringer, E. F., Kidd, J., Pramudianto, F., & Tinapple, D. (2016). Probing the landscape: Toward a systematic taxonomy of online peer assessment systems in education. In *Paper presented at the CSPRED 2016: Workshop on computer-supported peer review in education* (Raleigh, NC).

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*(5), 427–441.

Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin, 137*(6), 881–909.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220.

Binmore, K. G. (1994). *Game theory and the social contract: Just playing* (Vol. 2). MIT press.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education, 48*, 409–426.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11*(3), 468–481.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to inequality truly emotionally-driven? Dissociated behavioural and emotional responses in the ultimatum game task. *Cognition, 114*(1), 89–95.

Conybeare, J. A. C. (1984). Public goods, prisoners' dilemmas and the international political economy. *International Studies Quarterly, 28*(1), 5–22.

Corfman, K. P., & Lehmann, D. R. (1994). The prisoner's dilemma and the role of information in setting advertising budgets. *Journal of Advertising, 23*(2), 35–48.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*, 481–509.

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.

Fieseler, C., Bucher, E., & Hoffmann, C. P. (2019). Inequality by design? The perceived equality of digital labor on crowd working platforms. *Journal of Business Ethics, 156*(4), 987–1005.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, over-dispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392–404.

Gatfield, T. (1999). Examining student satisfaction with group projects and peer assessment. *Assessment & Evaluation in Higher Education, 24*(4), 365–377.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. New York; London: Longman.

Grogger, J. T., & Carson, R. T. (1991). Models for truncated counts. *Journal of Applied Econometrics, 6*(3), 225–238.

Hammer, T. H., & Landau, J. C. (1981). Methodological issues in the use of absence data. *Journal of Applied Psychology, 66*(5), 574–581.

Hayashi, N., Ostrom, E., Walker, J., & Yamagishi, T. (1999). Reciprocity, trust, and the sense of control: A cross-societal study. *Rationality and Society, 11*(1), 27–46.

Hayibor, S. (2017). Is fair treatment enough? Augmenting the equality-based perspective on stakeholder behaviour. *Journal of Business Ethics, 140*(1), 43–64.

Heerink, N., Mulatu, A., & Bulte, E. (2001). Income inequality and the environment: Aggregation bias in environmental Kuznets curves. *Ecological Economics, 38*(3), 359–367.

Hu, J., Blue, P. R., Yu, H., Gong, X., Xiang, Y., Jiang, C., & Zhou, X. (2016). Social status modulates the neural response to inequality. *Social Cognitive and Affective Neuroscience, 11*(1), 1–10.

Hughes, J. N. (2000). The essential role of theory in the science of treating children: Beyond empirically supported treatments. *Journal of School Psychology, 38*(4), 301–330.

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Journal of Instructional Science, 39*, 387–406.

Khandeparkar, K., Maheshwari, B., & Motiani, M. (2020). Why should I pay more? Testing the impact of contextual cues on perception of price inequality for the price-disadvantaged segment in dual pricing. *Tourism Management, 78*, 104075.

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior, 21*(6), 411–427.

Klein, G. (2018). The effectiveness of peer assessment and a proposal for its analysis using game theory. *The Journal of Education for Business, 93*(8), 436–442.

Kleine, M., Langenbach, P., & Zhurakhovska, L. (2016). Equality and persuasion: How stakeholder communication affects impartial decision making. *Economics Letters, 141*, 173–176.

Lacey, J., Carr-Cornish, S., Zhang, A., Eglinton, K., & Moffat, K. (2017). The art and science of community relations: Procedural equality at Newmont's Waihi Gold operations, New Zealand. *Resources Policy, 52*, 245–254.

Leigh, J. A., Signer, E. R., & Walker, G. C. (1985). Exopolysaccharide-deficient mutants of Rhizobium meliloti that form ineffective nodules. *Proceedings of the National Academy of Sciences, 82*(18), 6231–6235.

Lerche, T., & Kiel, E. (2018). Predicting student achievement in learning management systems by log data analysis. *Computers in Human Behavior, 89*, 367–372.

Li, H., Bialo, J. A., Xiong, Y., Hunter, C. V., & Guo, X. (2021). The effect of peer assessment on non-cognitive outcomes: A meta-analysis. *Applied Measurement in Education*, 1–25.

Lin, M. J. J., Hung, S. W., & Chen, C. J. (2009). Fostering the determinants of knowledge sharing in professional virtual communities. *Computers in Human Behavior, 25*(4), 929–939.

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

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211.

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

Li, A., Yan, Z., & Zhu, T. (2013). Self-report versus Web-log: Which one is better to predict personality of website users? *International Journal of Cyber Behavior, Psychology and Learning, 3*(4), 44–54.

Maddala, G. S. (1983). Methods of estimation for models of markets with bounded price variation. *International Economic Review, 24*(2), 361–378.

Mangelsdorf, K. (1992). Peer reviews in the ESL composition classroom: What do the students think? *ELT Journal, 46*(3), 274–285.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487–498.

McNamara, J. M., & Leimar, O. (2010). Variation and the response to variation as a basis for successful cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2627–2633.

Min, H. T. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing, 31*, 43–57.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.

Oishi, S., Kesebir, S., & Diener, E. (2011). Income inequality and happiness. *Psychological Science, 22*(9), 1095–1100.

Onwuegbuzie, A. J. (2012). Introduction: Putting the MIXED back into quantitative and qualitative research in educational research and beyond: Moving toward the radical middle. *International Journal of Multiple Research Approaches, 6*(3), 192–219.

Pandey, V., & Chatterjee, K. (2016). Game-theoretic models identify useful principles for peer collaboration in online learning platforms. In *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion* (pp. 365–368).

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*, 591–614.

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). 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.

Press, W. H., & Dyson, F. J. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences, 109*(26), 10409–10413.

Rae, D., Matlay, H., McGowan, P., & Penaluna, A. (2014). Freedom or prescription: The case for curriculum guidance in enterprise and entrepreneurship education. *Industry and Higher Education, 28*(6), 387–398.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education, 27*(3), 534–581.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45–78). Mahwah, NJ: Erlbaum.

Rodríguez-Gómez, G., Quesada-Serra, V., & Ibarra-Sáiz, M. S. (2016). Learning-oriented e-assessment: The effects of a training and guidance programme on lecturers' perceptions. *Assessment & Evaluation in Higher Education, 41*(1), 35–52.

Russell, A. A. (2004). Calibrated peer review: A writing and critical-thinking instructional tool. In *Invention and impact: Building excellence in undergraduate science, technology, engineering and mathematics (STEM) education* (Vols. 67–71)Washington, DC: American Association for the Advancement of Science.

Schunk, D. H., & Zimmerman, B. J. (2007). Influencing children's self-efficacy and self-regulation of reading and writing through modeling. *Reading & Writing Quarterly, 23*(1), 7–25.

Stewart, A. J., & Plotkin, J. B. (2013). From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proceedings of the National Academy of Sciences, 110*(38), 15348–15353.

Tangpong, C., Hung, K. T., & Ro, Y. K. (2010). The interaction effect of relational norms and agent cooperativeness on opportunism in buyer–supplier relationships. *Journal of Operations Management, 28*(5), 398–414.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 24–36.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–279.

Wang, Y., Li, H., Feng, Y., Jiang, Y., & Liu, Y. (2012). Assessment of programming language learning based on peer code review model: Implementation and experience report. *Computers & Education, 59*(2), 412–422.

Wang, Y., & Sun, F. (2018). How to choose an appropriate reviewer assignment strategy in peer assessment system? Considering equality and incentive. In *Proceedings of the 4th annual international conference on management, economics and social development (ICMESD 2018)* (Vol. 60, pp. 603–608). Atlantis Press.

Wu, W., Daskalakis, C., Kaashoek, N., Tzamos, C., & Weinberg, M. (2015). Game theory based peer grading mechanisms for MOOCs. In *Proceedings of the second ACM conference on learning@ scale* (pp. 281–286).

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 101826.

Xi, Y., Li, T., & Zheng, Y. (2013). Understanding cooperation in a single-trial Prisoner's Dilemma game: Interactions among three conditions. *Social Behavior and Personality: An International Journal, 41*(5), 721–729.

Yang, F., & Ott, H. K. (2016). What motivates the public? The power of social norms in driving public participation with organizations. *Public Relations Review, 42*(5), 832–842.

Yilmaz, K. (2013). Comparison of quantitative and qualitative research traditions: Epistemological, theoretical, and methodological differences. *European Journal of Education, 48*(2), 311–325.

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

Zou, Y., Schunn, C. D., Wang, Y., & Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assessment & Evaluation in Higher Education, 43*(5), 800–811.