Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment

Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, Scott Klemmer
UC San Diego Design Lab
La Jolla, CA 92093-0404 USA
{cmhicks, vipandey, cafraser, srk}@ucsd.edu

ABSTRACT
Peer assessment is rapidly growing in online learning, as it presents a method to address scalability challenges. However, research suggests that the benefits of peer review are obtained inconsistently. This paper explores why, introducing three ways that *framing* task goals significantly changes reviews. Three experiments manipulated features in the review environment. First, adding a numeric scale to open text reviews was found to elicit more explanatory, but lower quality reviews. Second, structuring a review task into short, chunked stages elicited more diverse feedback. Finally, showing reviewers a draft along with finished work elicited reviews that focused more on the work's goals than aesthetic details. These findings demonstrate the importance of carefully structuring online learning environments to ensure high quality peer reviews.

Author Keywords
Online education; peer review; massive open online courses; educational technology; distance learning

ACM Classification Keywords

K.3.1 [Computer Uses in Education]: Distance learning, Collaborative learning

INTRODUCTION: PEER ASSESSMENT POTENTIAL AND PROBLEMS

In online education, peer assessment has emerged as a crucial strategy for learner development [10, 15, 42]. Peer reviewers can provide social contact, feedback on work quality, and performance benchmarks for other students in online classes. Online platforms such as edX and Coursera frequently require learners to exchange peer reviews. Other models are being experimented with: for example, one platform recruits and pays former students as project reviewers [38]. Peer review software is available as both commercial (e.g., http://www.peerceptiv.com/) and open-source (e.g., http://www.peerstudio.org). The design of review environments is already producing real-world consequences for learners; these examples highlight the importance of understanding how different peer review features impact learners.

CHI'16, May 07-12, 2016, San Jose, CA, USA
© 2016 ACM. ISBN 978-1-4503-3362-7/16/05 $15.00
DOI: http://dx.doi.org/10.1145/2858036.2858195

As such, it is important to examine what peer review features result in high-quality feedback. Reviewing others' work can invoke powerful and unique benefits such as social motivation and learning [3, 10, 12]. However, research findings on peer reviewer efficacy are mixed, and difficult to reconcile. Some research has found that peer feedback can help online students improve at a rate comparable to instructor feedback [e.g. 9, 53]; other research points to limitations of peer review such as student distrust [8, 14, 29], grader inaccuracy [49], and inconsistency compared with expert reviewers [28, 44, 56, 57]. These results suggest that successfully scalable peer review requires strategies to ensure that high quality reviews are obtained more consistently. Evaluating peer assessment's potential is further complicated by the diversity of platforms and review structures examined.

This paper presents three ways that *framing* tasks in an online peer review environment can significantly change the reviews given. Three experiments manipulated features in an online peer review task to change the way the task was framed for peer reviewers. We investigated how changes to rubrics, task structure, and work representation impacted the quality, number of explanations and depth of feedback given by reviewers. We explored these changes across peer reviews for a variety of online artifacts (essays, resumés, and website designs). While not exhaustive, these findings may help strengthen the design of peer review environments. Table 1 summarizes these findings and practical recommendations for designing peer review tasks in online learning.

BACKGROUND RESEARCH

When designing peer assessment systems, it is useful to consider three factors that impact peer review quality: individual reviewer differences such as expertise, group differences such as cultural norms around assessment, and differences in the review processes.

Research on individual expertise and group differences can inform peer review. For example, novices and experts differ in the details they attend to: novices often overvalue detailed and surface-level concerns and undervalue conceptual and meaning-level issues. Novices may focus on grammar errors to the exclusion of revising logical flow in a writing

Finding	Recommendation
Numeric ratings in rubric increased explanations given in reviews but decreased overall review quality	Use objective scales like number or letter ratings for critical evaluation, but be aware that they may decrease developmental feedback
Short, separated stages for review increased goal-oriented feedback	Break up peer review tasks in stages to help broaden peer reviews' focus
Viewing a *sketched representation* shifted reviewers' focus to goals and purpose	Use artifacts that connect reviewers to the design and drafting process to deepen their feedback

Table 1: Summary of Findings and Recommendations for Framing Peer Review

assignment [59], and struggle to identify global problems [20, 25]. Expert reviewers may provide clearer justifications [58]. It also seems likely that individual differences in cognition and personality will impact peer reviewing: for example, reviewers with high executive functioning may give better feedback [35]. Group differences have also emerged for peer reviewers who belong to distinct populations, e.g: learners may tend to inflate assessments for peers from their own country [32].

Pivotal factors that impact peer reviews may also result from the immediate task *cues*. Decades of research have consistently found that norms and communication in classrooms can subtly impact learners' beliefs, with great ramifications for their behavior [1]. For example, the language and assessments used in classroom environments contain multiple cues for learners, supporting a range of beliefs about whether effort matters, whether self-disclosure is appropriate, and what mistakes mean [5, 11]. Learners often adopt these beliefs even when the messaging is implicit. Students who perceive course assessment to be harsh and unyielding will adopt an *avoidant* approach to learning that emphasizes escaping opportunities where one might fail [11], and these messages can be communicated through implicit cues, such as whom teachers call on.

How does the environment of an *online* assessment impact learner beliefs? And does this environment carry ramifications for peer review? These cues may be especially key in online environments, which often lack traditional communication cues, social presence, and immediate clarifications from instructors [e.g., 51]. In the absence of traditional communication cues such as body language, framing cues in an online setting may dramatically impact learners' task interpretation and subsequent performance.

FRAMING EFFECTS: THE IMPORTANCE OF REVIEW TASK FEATURES

Research on peer review content and quality has primarily focused on improving reviewers' performance relative to instructors or experts such as providing instructor training sessions for peer reviewers [e.g. 39, 47, 56]. Providing a rubric has also been found to increase reviewer performance [58]. However, along with such training, it is likely that tacitly implying assessment goals influences peer reviewers at all expertise levels. Without a careful understanding of these implications, peer review platforms may inadvertently create *framing effects* which change the focus and content of peer reviews.

Framing effects have been shown to significantly affect people's behavior [e.g., 9, 18, 22, 34, 40]. While framing effects are complex and multifaceted, one salient effect for peer reviews is anchoring and adjustment, a heuristical behavior where an initial "anchor point" can inordinately influence an individual's later estimations and behaviors. If the initial anchor is high or low, later estimations will also be high or low respectively [19, 52]. For example, assessments of university reputation are influenced by known prior assessments [4].

This paper investigates the following questions: How might such framing cues present in peer review? Do unspoken and implicit design signals impact peer reviewers' conception of the review process?

EXPERIMENT 1: GIVING REVIEWERS NUMERIC RATINGS PROMPTS MORE EXPLANATION

Classroom research has uncovered significant effects on learner disclosure and perseverance based on whether students believe that educational goals are *critical* or *developmental* [11]. For online classes, this belief could also be an important factor in peer review. A reviewer who perceives their task to be assigning a final grade may give critical feedback that focuses on identifying flaws. By contrast, a reviewer who perceives their task to be helping a learner improve may provide more explanations and suggest changes to the reviewee.

One common cue that may affect whether feedback is seen as critical or developmental is asking reviewers to assign numeric ratings. Online, review rubrics vary widely, and may or may not ask reviewers to rate work on an explicit scale. For example, Figure 1 shows a review rubric used in a recent Coursera course on Cognitive Ethnography which did not use numeric ratings [36]. In contrast, Figure 2 shows a rubric from a recent edX course, The Art of Poetry, which asked learners to review peers' essays with multiple numeric point scales [43]. One possibility is that numeric ratings draw reviewers' focus to fault-finding, creating more negative feedback in their text comments. Alternately, numeric ratings may remove the perceived responsibility of providing explicit ranking, and encourage them to focus instead on developmental suggestions in their text feedback.

Experiment 1 provides some suggestive work exploring this question.

METHOD

Experiment 1 examined whether providing numeric ratings changes the content of peer reviews for graduate application essays.

Participants

Participants were recruited from online advertisements and flyers on a Southern California university campus. 53 graduate school applicants participated in this experiment (see Table 2). All participants had completed a graduate application essay, and were actively applying for graduate school within the time period of the study. We used these selection criteria to study peer reviewers for whom the review task would be relevant. By using a peer review task outside of an existing online course, we were able to examine the effects of peer review on a task for which peers had received no prior influence or training from a course rubric or instructor.

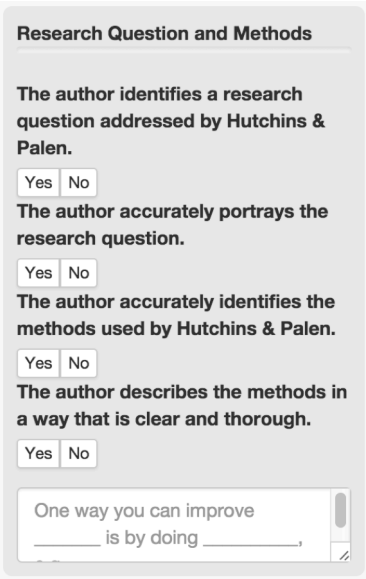

Figure 1: Online Peer Review Rubric: No numeric ratings required of peer reviewers [36]

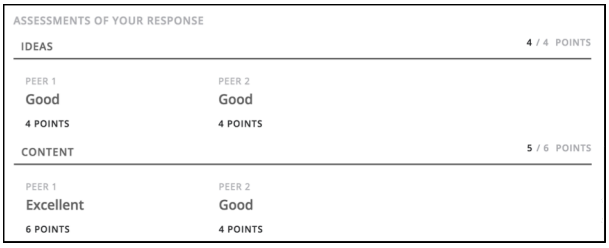

Figure 2: Online Peer Review Rubric: Numeric ratings required from peer reviewers [43]

Materials and Procedure

Participants enrolled through an online website created by the authors, and provided their area of application, college enrollment, and whether they were native English speakers. Native and non-native (ESL) English speakers were assigned in equal numbers to the two conditions in order to control for language confounds across conditions.

Participants uploaded their graduate application essay for peer review through PeerStudio, an online peer review platform [31, 33]. Before completing the review task, participants were directed to further resources for graduate application essay writing from a university careers website [54]. After submitting their essay, participants were required to review two other essays before receiving feedback on their own. Participants were then asked to submit revised drafts for a second round of feedback. The study was open to enrollment and participation for one month, and participants were allowed to work at their own pace during that time.

Experimental Conditions

Participants were randomly assigned to receive and give reviews in one of two conditions: **numeric** (28 participants) or **non-numeric** (27 participants). *Both* conditions provided reviewers with open text fields for comments on the essay's thesis statement, supporting evidence, and conclusion. In the numeric condition, reviewers were also asked to rate the thesis statement, supporting evidence and conclusion respectively on a scale from 1 to 5 (Figure 3).

RESULTS

53 participants submitted at least one essay; 24 submitted a revised essay after receiving feedback. 12 participants submitted more than 2 drafts, with one submitting 12 drafts. In total, 205 reviews were submitted: 115 in the numeric condition and 90 in the non-numeric condition. Because some reviewers failed to complete all the tasks, the number of observations varies in the following analyses.

Analysis Plan

Text responses were numerically coded by three raters who were blind to the experimental conditions. Table 3 shows a complete list of the measures coded. The primary depend-

Age	*M* = 25.11	Range = 22 - 30
Gender	Female = 22	Male = 31
Graduate Application Areas	Science = 70%	Social Science = 21%
	Arts & Humanities = 9%	
Enrolled in College	Yes = 41%	No = 59%
ESL	Yes = 12%	No = 88%
Ethnicity	Caucasian = 24%	Asian = 14%
	Hispanic = 6%	Chose not to answer = 56%

Table 2: Descriptive Statistics for Experiment 1

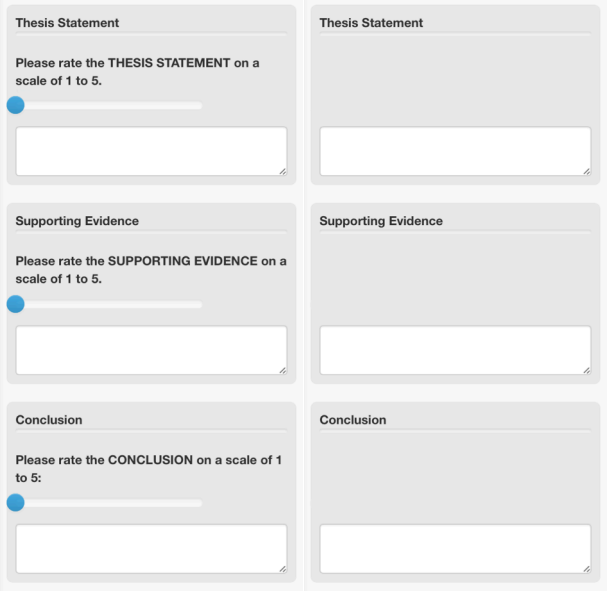

Figure 3: Numeric (left) and non-numeric (right) feedback conditions in Experiment 1. Peer reviewers in the numeric condition were provided with both open-ended text boxes and numeric ratings.

ent variables for our analyses were the number of explanations, the number of positive and negative comments, and ratings of overall quality. Explanation received one point if a comment had a suggestion for improvement and justified it with an explicit explanation (e.g. from participant comments, "change the conclusion because it is repetitive and will be boring"). No explanation count was given if a comment had no suggestions or made only vague statements (e.g., "change the conclusion"). Negative comments (e.g., "I hate your intro") were counted only when the reviewer made a purely negative statement with no explicit suggestion for change. "Quality" was a subjective 1-5 rating for how helpful raters judged a review to be in comparison to other reviews. Ratings were positively correlated (r = .82-.85, p < .0001) and the average from all ratings was used for analysis.

Quantitative Results

The presence of numeric ratings had a significant effect on the content of reviews (Figure 4): reviewers in the numeric condition gave *more* explanations (M = 2.60) than reviewers in the non-numeric condition (M = 2.03), $F(1, 120)$ = 4.04, p = .04. This effect size was moderate (Cohen's effect size d = .45). Reviewers in the numeric condition were also significantly more likely to make positive comments (M = 1.75 vs. M = 1.02), $F(1, 120)$ = 8.23, p < .01. This effect size was moderate to large (Cohen's d = .63).

However, reviews in the non-numeric condition were rated higher *quality* (M = 2.80 versus M = 2.36), $F(1, 120)$ = 4.05, p < .05. This effect size was moderate (Cohen's d = .44). Interestingly, reviewers in the non-numeric condition wrote significantly *longer* reviews (M = 336) than those in the numeric condition (M = 126), $F(1, 120)$ = 6.42, p = .01. This effect size was moderate (Cohen's d = .57).

Explanation	Number of explanations for improvement
Positive Comments	Number of non-critical positive comments (e.g., "I loved the examples from your lab research")
Negative Comments	Number of critical negative comments (e.g., "your conclusion was terrible")
Length	Total character count for each submitted review
Quality	Quality score for a review rated on a 1-5 scale

Table 3: Dependent measures used in text analysis for Experiment 1

Explanation and Positive Comments were not correlated. There was no significant difference in Negative Comments between conditions, however, Explanation was positively correlated with Negative Comments, $r(121)$ = .60, p < .0001.

Preliminary analyses found no effect of Gender, Age, Enrollment in College, Ethnicity or ESL, therefore these variables were excluded from subsequent analyses. The number of revisions submitted also did not vary significantly between conditions.

Experiment 1 Discussion

Providing a numeric rating scale along with the open-ended review prompted more explanations and more positive comments. However, open-ended reviews with no such scale were rated more helpful as feedback for the writer.

One possible explanation for these findings is that a numeric scale prompts reviewers to think their task is primarily *making the assessment standard clear* to the reviewee: re-

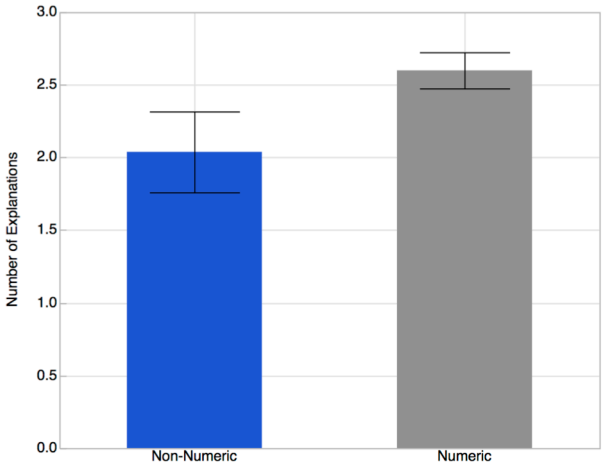

Figure 4: Explanation Feedback is higher in numeric condition in Experiment 1. Error bars show standard error.

quiring a rating emphasized the reviewer's explanatory role. While reviewers in the numeric condition provided more explanations, their reviews appeared to focus more on explaining grading and less on providing help to the reviewee. Their more positive tone overall may indicate that reviewers in the numeric condition felt the need to justify or console the reviewee for their grade. Without the numeric cue, reviewers provided feedback that was rated as more helpful, perhaps because their interpretation of the task remained developmentally focused. Future research that includes post-task interviews with reviewers could help to clarify the motivations behind these observed differences.

Experiment 1 suggests that rubric features can produce significant changes in review content, and possibly impact reviewers' conceptions of their task goals. Experiment 2 explores whether similar effects are found for the structure of the review task.

EXPERIMENT 2: DIVIDING REVIEW INTO SHORTER TASKS PROMPTS GOAL-ORIENTED FEEDBACK

Design choices made about the structure of tasks in online learning environments can have significant consequences. For example, many Massive Open Online Courses (MOOCs) present short (e.g., less than 10 minutes) lecture videos interspersed with multiple choice questions rather than a single, hour-long lecture typical of universities. This structural choice stems from the insight that shorter content consumption followed by immediate interaction is often more engaging and beneficial for learners [e.g., 30]. The principle of *active learning* helps to explain this finding, arguing that immediately engaging students in activities reinforces novel material, thereby prompting deeper learning and longer retention of material [e.g., 45, 46].

Can task structures also impact *reviewers'* behavior? It is possible that reviewers will generate better reviews when their task is divided into shorter discrete steps rather than a single, longer task because this prompts more active engagement with the review task. It is also possible that framing the review task with short steps may cue reviewers to consider more diverse types of feedback rather than perseverating on a single feature in the reviewed work. Novice reviewers tend to overdwell on surface details and fail to recognize global problems in work [20, 25, 59]. Will framing a review activity as multiple small review tasks prompt reviewers to broaden the scope of their reviews?

On the other hand, it is also possible that dividing a review task will produce sparser and lower quality reviews. Task interruptions can be highly disruptive to performance, introducing a mental load which decreases both performance quality and time on task [24, 50]. Reviewers may generate longer reviews, and note more features, when they are given uninterrupted time on task. It is also possible that dividing the task will produce no difference: reviewers may generate a certain amount of feedback regardless of the review task structure.

Experiment 2 tests whether a different structure for the review process changes review content. We hypothesized that breaking up a review task into a series of shorter critiques improves reviewers' attention to overarching goals of the work. As in Experiment 1, we chose to test peer reviewers outside of the context of an online course. We chose resumé review as the experimental task: resumé writing is a familiar task for most people who have searched for a job, and a real-world experience for participants in this experiment.

METHOD

Experiment 2 examined whether reviewers giving feedback in shortened stages would give different reviews compared with reviewers completing the task in one long session.

Participants

Participants were recruited from an experimental subject pool at a Southern California university (see Table 4). All had completed a professional resumé, and were either applying for jobs or planning to apply for jobs within the next six months. These criteria selected peer reviewers for whom resumé review would be relevant. Participants received undergraduate course credit for participation in this experiment. Participants also self-reported their perceived academic standing, compared with classmates, on a 1-5 scale.

Materials and Procedure

The resumés in this experiment were created by the researchers in collaboration with undergraduate research assistants. Resumés with a range of believable work experience were created for fictional undergraduate Cognitive Science majors.

The experiment was administered online. Participants completed a short, three-question training that familiarized them with some best practices by asking them to review their own resumés. This training helped ensure that participants had at least some preliminary knowledge of resumé improvement, and encouraged them to see review as relevant to their personal experience. This training led participants through the resumé features of layout, language, and content, and how those features supported professional goals (e.g., clarity of language may help a recruiter to perceive the writer as thoughtful and intelligent).

Age	*M* = 20.69	Range = 18 - 27
Gender	Female = 60	Male = 24
Ethnicity	Asian = 51%	Latino = 19%
	Caucasian= 20%	Other = 10%
Major	Psychology = 48%	Biology = 34%
	Other Social Science = 7%	Other = 11%
ESL	Yes = 17%	No = 83%

Table 4: Descriptive Statistics for Experiment 2

Experimental Conditions

Participants were randomly assigned to one of two conditions: structured review (43 participants) or unstructured (41 participants):

1. In the **Structured** condition, participants were asked to give feedback on an undergraduate student resumé in three steps, with each step titled as a different resumé feature: layout, language, and content. Steps were titled with general resumé characteristics to provide participants with some explanation for the discrete steps. Participants navigated through three pages, with a small open text field displayed on the same page as the resumé.
2. In the **Unstructured** condition, participants were shown the same student resumé. They were asked to give feedback in one step, with a single, large open text field displayed on the same page as the resumé.

RESULTS

84 participants provided feedback on resumés.

Analysis Plan

Text responses from participants were coded by three research assistants who were blind to conditions. Ratings were positively correlated (r = .8-.86, p < .0001); analysis used the average rating for each measure. Table 5 shows a complete list of the measures coded. The primary dependent variables for analysis were the review comments on aesthetic features and writer goals, and ratings of overall feedback quality in comparison to the other reviews. Quality was scored by answering the prompt, "if this were feedback on your work, how helpful would you find it?" For all ratings where a discrepancy arose, raters reached an agreement on the score after discussion.

Aesthetic Feedback	Number of statements about aesthetic details of resumé such as font, bullet styles, or layout
Goal Feedback	Number of statements specifically mentioning reaching the resumé writer's goals such as significant rewording to present a specific professional identity
Positive Comments	Number of non-critical positive comments (e.g., "I really liked your formatting")
Negative Comments	Number of critical negative comments (e.g., "The large bullet points really irked me")
Explanations	Number of explanations given for suggested changes (e.g., "change the layout *because it's hard to read*")
Quality	Quality score for a reviewer rated on a 1-5 scale by three raters: asked "how helpful would you find this review, if this were your resumé?"
Length	Total character count for each submitted review

Table 5: Dependent measures used in text analysis for Experiment 2

Quantitative Results

Structured review produced significantly more feedback on writer goals, $F(1,83) = 27.83$, $p < .0001$ (Figure 5). This effect size was large (Cohen's effect size $d = 1.16$). In addition to commenting more on global, goal-driven features of resumés, reviews from the structured condition were rated as significantly higher quality than reviews from the control condition, $F(1,83) = 11.64$, $p = .001$. This effect size was moderate to large (Cohen's $d = 0.71$). Structured review also produced significantly more positive comments throughout the experiment, $F(1, 83) = 44.35$, $p < .0001$, with a large effect size (Cohen's $d = 2.08$). There was no statistically significant difference in the number of negative comments given between conditions.

There was no significant difference in the number of feedback suggestions participants made for aesthetic features of the resumés. Across both conditions, the amount of Goal Feedback was correlated with review quality, $r(84) = .64$, $p < .0001$. Reviews in the structured condition were also longer than in the unstructured condition ($M = 693$ compared with $M = 465$); $F(1, 83) = 12.29$, $p < .001$, with a medium to large effect size (Cohen's $d = 0.77$).

Additionally, for the 68 participants who volunteered this information, self-reported academic rank was positively correlated with review quality, $r(68) = .35$, $p = .003$. This provides some confirmation of the Quality ratings, as reviewers who gave high quality reviews were also better students.

Experiment 2 Discussion

In Experiment 2, providing discrete stages in the review task improved the quality and broadened the focus of reviewers' feedback. While all reviewers gave feedback on aesthetic choices like font and layout, the structured review condition prompted reviewers to attend more to the submis-

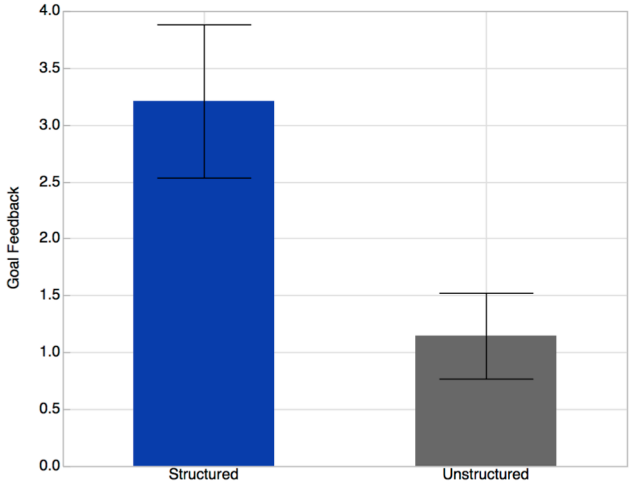

Figure 5: Review Comments on Writer Goals are significantly higher for Structured Condition in Experiment 2.

sion's goals and underlying message. The raters' "Quality" scores correlated with the amount of feedback reviewers gave on work goals, suggesting that raters interpreted reviews with goal feedback as being more helpful. Future work should test whether feedback recipients themselves perceive goal feedback as more helpful.

Experiment 2 suggests that choices about the way a peer review task is structured can significantly impact reviews. Experiment 3 examines how choices about the representation of the reviewed work itself can impact reviews. Experiment 3 also examined this impact with a subject pool of Mechanical Turk workers. MTurkers provide demographic characteristics which are broader than the subject pool used in Experiment 2, allowing a closer comparison to the general online learning populations (e.g., Coursera reports that average age for users in many courses is 35 [37]). However, it is also important to note that there may be differences between these populations which impact the generalizability of these findings. For instance, MTurkers are likely more experienced in ratings tasks than an average student participant.

EXPERIMENT 3: SHOWING DRAFTS ENCOURAGES REVIEWERS TO FOCUS ON PROCESS

Choices about how work is *represented* may also frame peer review tasks for learners. Often, academic performance is evaluated on a static end product: in a traditional classroom, a final paper, project, or design is submitted for review and evaluation. However, growing emphasis on process and showing drafts in educational research [e.g., 2, 9, 48] raises interesting possibilities for online peer review. Does seeing a draft example of work change the kind of feedback that peers give? And does the way that draft is represented affect the kinds of feedback that peers generate?

One theoretical explanation for why seeing drafts could elicit better performance is that drafts encourage *growth mindset* beliefs [17]. A large body of educational research has found that emphasizing the process of work versus the static outcome leads learners to more highly value their own effort, and engage more deeply and persistently with the work [e.g., 7, 26, 41]. For example, children who were given *process praise* ("you must have tried really hard to do this") instead of *outcome praise* ("that's the right way to do it") showed both more persistence and more adaptive beliefs about learning [27]. Some intriguing anecdotal accounts also highlight the importance of draft representation specifically: for example, Buxton recounts how architects intentionally show lower quality, more "hand-drawn" sketches to clients in order to shift their focus to global design questions and away from surface details [6].

How might representation framing impact giving feedback to creators? It could be that seeing a draft provides a growth mindset frame, encouraging reviewers' deepened engagement in the process. On the other hand, this framing may not be strong enough to impact reviewers; in a usability study, no significant difference was found for reviewers' identification of usability issues across sketched, wireframe, or fully rendered website designs [55]. This question is particularly relevant for the evaluation of visual, creative work. For example, Coursera's Interaction Design Specialization [13] requires learners to evaluate and generate feedback on creative visual work such as interface prototype materials.

Experiment 3 examined whether the way a website design is represented to reviewers will impact the reviews given. We hypothesized that seeing different representations of an evaluation piece can frame the task goals for reviewers: seeing unfinished or "rough" versions of work could prompt reviewers to consider the creator's process and goals for the work product. Additionally, Experiment 3 contrasted this frame with explicitly telling reviewers about the website creator's goals, asking whether this would similarly impact review.

METHOD

Experiment 3 examined whether providing additional representation that demonstrated the work process behind an artifact would change the focus of reviews.

Participants

Participants were recruited from Amazon Mechanical Turk, https://www.mturk.com, and compensated $2 for completing an approximately twenty-minute experimental exercise (see Table 6).

Materials and Procedure

Based on pilot testing, participants were given 20 minutes to complete the experimental task. Participants were randomly assigned to one of three conditions: sketch interface, wireframe interface, or a full-fidelity condition. Participants were also randomly assigned to one of two exploratory conditions: context or no context for the assignment goals. Three website designs were created to represent a fictional oceanic conservation organization. The designs created were intended to represent a believable student project in early draft form.

Age	*M* = 34.12	Range = 19 - 65
Gender	Female = 104	Male = 115
Ethnicity	Caucasian = 70% Latino = 7% Other = 5%	Asian = 12% African American = 6%
Web Design Expertise	Beginner = 66% Intermediate = 31%	Expert = 3%

Table 6: Descriptive Statistics for Experiment 3

Experimental Conditions

All participants completed a short, three-question pre-test training that led them through features of website design. Participants were then asked to give feedback on a student's website design.

Experiment 3 employed a 2x3 design. Firstly, participants were randomly assigned to one of two exploratory conditions:

1. In the **Context Condition**, participants were told: "*The student designer you've been assigned to review has submitted the following description of their assignment goals: 'To design an ecology website that will make a user interested in learning more about the ocean, and/or help people get involved in conservation.'*"
2. In the **No Context Condition,** participants were simply told that the student designer had submitted the following website design for review, with no contextual information about the goals of the website.

All participants saw the finished website design (Figure 8), and were given time to observe and make notes on this design. Participants were then randomly assigned to one of three **experimental conditions** while writing their reviews:

a. *Sketch condition*. Participants in the Sketch condition saw a simple hand-drawn sketch of the website design (Figure 6).

b. *Wireframe condition*. Participants in the Wireframe condition saw a simple wireframe mockup of the website design (Figure 7).

c. *Full-fidelity condition.* Participants in the Full-fidelity condition simply saw the finished website design again (Figure 8).

RESULTS

Analysis Plan

219 participants completed this experiment (see Table 6). Text responses were coded by three raters who were blind to the experimental conditions. Ratings were positively correlated (r = .85-.89, $p < .0001$) and the average rating for each measure was used for analysis. Table 7 lists the measures coded. The primary dependent variables for our analyses counted the amount of feedback reviewers gave on aesthetic features of the website, and the amount of feedback reviewers gave on the user process and experience. "Process" feedback concerned overall goals for the website design, for instance, feedback on changes that might keep a user engaged and interested in the organization's story. "Aesthetic" feedback, on the other hand, concerned purely visual features of the website design such as the colors chosen.

Quantitative Results

Reviewers who saw the Sketch representation gave significantly more process feedback (M = .61) compared with the Wireframe (M = .41) and the Full-fidelity condition (M =

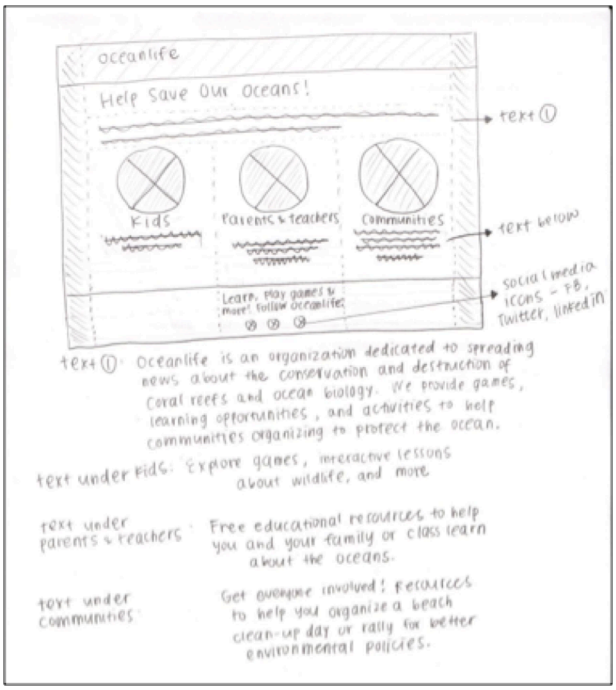

Figure 6: Sketch graphic for Experiment 3

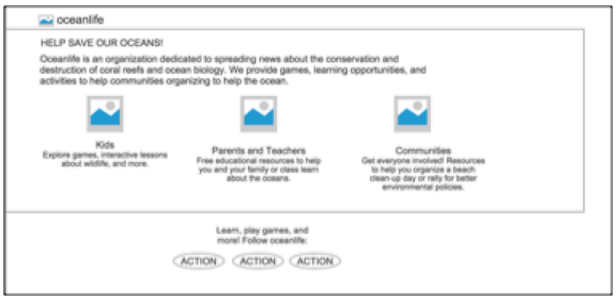

Figure 7: Wireframe graphic for Experiment 3

Figure 8: Website graphic for Experiment 3

.34), $F(2, 216) = 4.14$, $p = .01$ (Figure 9). This effect size was moderate (Cohen's effect size $d = .45$). As in Experiment 2, there was no statistically significant difference in the amount of feedback given on aesthetic features between conditions: while reviewers in the sketch condition gave more process feedback, they did not give *less* feedback on aesthetics.

Reviews in the Sketch condition were also significantly longer than the other two conditions ($M = 210$ compared with $M = 162$ for the Wireframe condition and $M = 160$ for the Full-fidelity condition); $F(2, 216) = 5.86$, $p = .003$. This effect size was moderate (Cohen's effect size $d = .50$).

Finally, reviewers in the Sketch condition remarked on a greater overall number of features ($M = 1.93$) compared with the Wireframe ($M = 1.52$) and Full-fidelity conditions ($M = 1.82$); $F(2, 216) = 3.86$, $p = .02$. This effect size was moderate (Cohen's $d = .32$). The number of explanations given, amount of positive or negative comments, and the number of specific recommendations for improvement, did not differ between conditions.

Context Condition

No significant effects were found for the Context condition: participants who were shown the additional context emphasizing the website goals did not produce measurably different reviews than participants without this information.

Demographic Effects

Of the descriptive variables, Age, Ethnicity, and Gender, only Gender had a significant effect on the number of explanations given. Female participants tended to give more explanations ($M = .70$) than male participants ($M = .42$), $F(1, 215) = 4.76$, $p = .03$; this effect size was small to moderate (Cohen's effect size $d = .36$). Gender was therefore included in subsequent analyses but no further main or interaction effects were found. This effect is interesting, but preliminary; no gender differences were found in the previous experiments. No effects were found for self-reported website design expertise.

Aesthetic Feedback	Number of statements on font, color, or layout (e.g., "Change the font")
Process Feedback	Number of statements on broader user process and experience (e.g., "You need to make it clearer how people can engage with this organization, so they will get excited")
Total Number of Features	Number of distinct features the reviewer commented on
Explanations	Number of explanations reviewers gave for their feedback (e.g., "Change the font *because it is hard to read like that*")
Recommendations	Number of concrete suggestions reviewers gave within their feedback (e.g., "Change the font, *you could make it dark gray instead*")
Negative Comments	Number of negative statements about the design (e.g., "I really hate this layout")
Positive Comments	Number of explicit statements of personal positive opinion towards the design (e.g., "I loved the pictures chosen!"
Length	Character count for each submitted review

Table 7: Dependent measures used in text analysis for Experiment 3

DISCUSSION

In this paper, three experiments illustrated how the framing of peer review can significantly impact reviews. Feature changes in rubrics, task structure, and artifact representation resulted in reviews that were significantly different in both the quality and focus of reviewer feedback.

Experiment 1 found that providing a numeric scale with reviews elicited more explanations from reviewers, but open reviews with no scale were higher in quality. One possible interpretation of these results is that the presence of a rating scale emphasized the reviewer's *explanatory* role, but decreased the *developmental* role. It may be that reviewers took their role to be both explanatory, and consolatory: numeric reviewers also gave more positive affirmations. Previous research has found that positive reviews are received more favorably [58], and it could be that reviewers in the numeric condition use more complimentary language as a social tactic to diminish the negative impact of delivering a grade. Novice reviewers often underexplain their comments [23, 57], and feedback with explained suggestions for change leads students to make more improvements [21]. We hypothesize that peer review systems could benefit from design that elicits more explanations. It is possible that novice reviewers underexplain their feedback without being

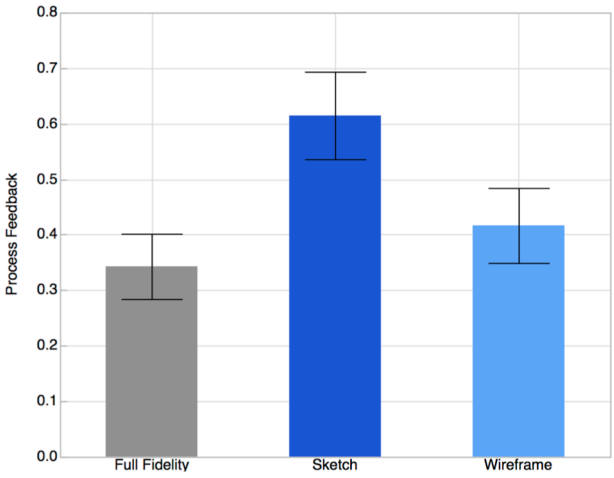

Figure 9: Process Feedback is higher in Sketch Condition by Condition in Experiment 3

aware of this, and including a numeric scale may help peer assessments that require extra clarity, such as final graded assignments. Nevertheless an important *benefit* of requiring open-ended feedback could be that this structure emphasizes developmental over critical feedback. Peer review without a numeric scale may therefore be helpful for peer assessments with developmental goals, such as early assignments or works in progress. Future research should directly test these hypotheses, and qualitative data such as interviews could help describe how these design choices impact reviewers' mindsets.

Experiment 2 found that discrete task sections elicited reviews that were rated more helpful and that included feedback on *both* the writer's career goals and aesthetic features. One interpretation for this finding is that reviewers given an unstructured task perseverate on the resumés' surface features, as novice reviewers fixate on grammar instead of essay structure [59]. In contrast, a divided structure task may have highlighted the need to generate comments that are different from the feedback already given. When asked to re-start a task, reviewers have an opportunity to reconsider the task's aims. This task structure may also encourage active learning, maintaining the reviewer's engagement by presenting shorter, repeated interactions instead of a single, long interaction at the end of a longer task [45]. It is difficult to sustain learner attention and engagement in online learning environments [33], particularly for tasks that require time and cognitive effort. Peer assessment could benefit from using smaller repeated tasks to maintain reviewer engagement. However, it is also possible that repeated tasks result in task interruption, carrying trade-offs for reviews' depth or comprehensiveness. Future research should investigate how task structure can help to support engagement and provide scaffolding for high effort learning activities in online environments, as well as test the theoretical explanations of these benefits.

Experiment 3 found that showing a sketch of a design during reviews elicited significantly more feedback on user experience and the global goals for the work. Although all participants were instructed to give feedback on the final design, hand-drawn representations may have cued reviewers to consider the dynamic possibilities behind a static outcome. This cue may encourage reviewers to invest more deeply in helping to change the work, viewing it as a malleable process rather than a fixed product [16, 17, 48]. It is also interesting that this effect was not found for a wireframe draft of the same design. It is possible that the sketch was more readily interpreted as a "draft" by reviewers and that a wireframe was interpreted as a more finished product, but future research should explore why different kinds of drafts do or do not produce similar effects. Intriguingly, giving reviewers explicit information about the reviewee's user goals also did not significantly change the content of their reviews. This preliminary result suggests that framing features could influence reviewing even when peer reviewers are unresponsive to explicit training that attempts to shape their reviews. Future research should explore the relationship between explicit and implicit framings.

Experiment 3 explored framing effects with Mturk workers, who are more similar to online learning populations in some characteristics (e.g., in age and ethnicity), but less similar in others such as compensation and experience with ratings tasks. This paper examined framing effects *within* these diverse populations in order to explore general principles that can help design online review systems, but comparing *between* such populations is a key question for future work. Further research is needed to examine whether framing design choices produce differing effects between diverse reviewing environments (e.g., reviewing commercial work compared with work in open online courses). Examining creators' reactions to the feedback on their work is also an important next step, and future research should incorporate reviewees' reactions into the evaluation of review quality as well as test whether students' academic outcomes in real-life review environments correspond to reviews' rated helpfulness.

Understanding framing effects is particularly compelling for online learning platforms as they aim to scale for global audiences. In order to achieve such scale, these learning environments will often need to function algorithmically and independently of instructor oversight. In such a design, framing effects may be amplified across multiple iterations and learners. Intentionally framing educational environments in ways that elicit high quality performance can be a powerful lever to improve learning.

ACKNOWLEDGMENTS

We thank Rachel Chen & Crystal Kwok for help conducting and analyzing this research.

REFERENCES

1. Carole Ames. 1992. Classrooms: Goals, structures, and student motivation. *Journal of Ed Psych* 84, 3, 261-271.
2. Erik Borg and Mary Deane. 2011. Measuring the outcomes of individualised writing instruction: A multilayered approach to capturing changes in students' texts. *Tech in High Educ*. 16, 3, 319-331.
3. David Boud, Ruth Cohen, and Jane Sampson (Eds.). 2014. *Peer learning in higher education: Learning from and with each other*. Routledge.
4. Nicholas A. Bowman and Michael N. Bastedo. 2011. Anchoring effects in world university rankings: Exploring biases in reputation scores. *High Educ*. 61, 4, 431-444.
5. Paul C. Burnett. 2002. Teacher praise and feedback and students' perceptions of the classroom environment. *Educ Psych*. 22, 1, 5-16.

6. Bill Buxton. 2010. *Sketching user experiences: Getting the design right and the right design.* Morgan Kaufmann.
7. Karen Chalk, and Lewis A. Bizo. 2004. Specific praise improves on-task behaviour and numeracy enjoyment: A study of year four pupils engaged in the numeracy hour. *Educ Psych in Prac*. 20, 4, 335-351.
8. Winnie Cheng and Martin Warren. 1997. Having second thoughts: Student perceptions before and after a peer assessment exercise. *Stud in High Educ*. 22, 2, 233-239.
9. Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learn and Instruc*. 20, 4: 328-338.
10. Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Comput. Educ*. 48, 3 (April 2007), 409-426.
11. Marcy A. Church, Andrew J. Elliot, and Shelly L. Gable. 2001. Perceptions of classroom environment, achievement goals, and achievement outcomes. *Jour of Educ Psych*. 93, 1, 43-54.
12. D. Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A. Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1139-1152. http://dx.doi.org/10.1145/2675133.2675251
13. Coursera. Interaction Design Specialization [website]. Retrieved September 23, 2015 from https://www.coursera.org/specializations/interaction-design/
14. Phil Davies. 2000. Computerized peer assessment. *Inno in Educ and Teach Inter.* 37, 4, 346-355.
15. Elaine DiGiovanni and Girija Nagaswami. 2001. Online peer review: An alternative to face-to-face? *ELT Jour*. 55, 3, 263-272.
16. Carol S. Dweck. 1986. Motivational processes affecting learning. *Am Psych*. 41, 10, 1040-1048.
17. Carol S. Dweck. 2006. *Mindset: The new psychology of success.* Random House.
18. Carol S. Dweck and Ellen L. Leggert. 1988. A social-cognitive approach to motivation and personality. *Psych Rev*. 95, 2, 256–273.
19. Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psych Sci*. 17, 4, 311-318.
20. Nancy Falchikov and Judy Goldfinch. 2000. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Rev of Ed Res*. 70, 3, 287-322.
21. Dana R. Ferris. 1997. The influence of teacher commentary on student revision. *TESOL Quarterly.* 31, 2, 315-339.
22. Kristel M. Gallagher and John A. Updegraff. 2012. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. *Annals of Beh Med*. 43, 1, 101-116.
23. Joo Seng Mark Gan. 2011. *The Effects of Prompts and Explicit Coaching on Peer Feedback Quality.* Doctoral Dissertation. The University of Auckland, Auckland, New Zealand.
24. Tony Gillie and Donald Broadbent. 1989. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psych Res* 50, 4, 243-250.
25. John R. Hayes, Linda Flower, Karen A. Schriver, James F. Stratman, and Linda Carey. 1987. Cognitive processes in revision. In *Advances in applied psycholinguistics, Volume II: Reading, writing, and language processing*, Sheldon Rosenberg (Ed.). Cambridge, England, Cambridge University Press, 176-240.
26. Jennifer Henderlong and Mark R. Lepper. 2002. The effects of praise on children's intrinsic motivation: A review and synthesis. *Psych Bull.* 128, 5, 774-795.
27. Melissa L. Kamins and Carol S. Dweck. 1999. Person versus process praise and criticism: Implications for contingent self-worth and coping. *Dev Psych*. 35, 3, 835-847.
28. James C. Kaufman, John Baer, Jason C. Cole and Janel D. Sexton. 2008. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creat Res Jour*. 20, 2, 171-178.
29. Julia H. Kaufman and Christian D. Schunn. 2011. Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instruc Sci*. 39, 3, 387-406.
30. Juho Kim, Elena L. Glassman, Andrés Monroy-Hernández and Meredith Ringel Morris. 2015. RIMES: Embedding interactive multimedia exercises in lecture videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 1535-1544. http://dx.doi.org/10.1145/2702123.2702186
31. Chinmay Kulkarni, Michael S. Bernstein and Scott Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings from The Second (2015) ACM Conference on Learning@ Scale (L@S '15)*. ACM, New York, NY, USA, 75-84. http://dx.doi.org/10.1145/2724660.2724670
32. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller and Scott R. Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Trans. Com-*

put.- Hum. Interact. 20, 6, Article 33 (December 2013), 31 pages. http://dx.doi.org/10.1145/2505057

33. Yasmine Kotturi, Chinmay Kulkarni, Michael S. Bernstein and Scott R. Klemmer. 2015. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale (L@S '15)*. ACM, New York, NY, USA, 31-38. http://dx.doi.org/10.1145/2724660.2724676
34. Irwin P. Levin, Gary J. Gaeth, Judy Schreiber and Marco Lauriola. 2002. A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Org Beha and Hum Dec Proc*. 88, 1, 411-429.
35. S. S. J. Lin, E. Z. F. Liu and S. M. Yuan. 2001. Web-based peer assessment: Feedback for students with various thinking-styles. *Jour of Comp Assi Learn*. 17, 4, 420-432.
36. Melanie McComsey. 2015. 102B: Cognitive Ethnography. (Jan-Mar 2015). University of California, San Diego. Retrieved from https://www.peerstudio.org/courses/19
37. Katie McEwan. 2013. Getting to know Coursera: Who is everyone? (20 February 2013). Retrieved November 30, 2015 from https://cft.vanderbilt.edu/2013/02/getting-to-know-coursera-who-is-everyone/
38. Medium. How a Udacity graduate earns $11k a month reviewing code. Retrieved September 21, 2014 from http://www.medium.com/@olivercameron/how-a-udacity- graduate-earns-11k-a-month-reviewing-code-c2a7d295724c
39. Hui-Tzu Min. 2005. Training students to become successful peer reviewers. *System.* 33, 2, 293-308.
40. Thomas E. Nelson, Zoe M. Oxley and Rosalee A. Clawson. 1997. Toward a psychology of framing effects. *Pol Beh*. 19, 3, 221-246.
41. Eleanor O'Rourke, Kyla Haimovitz, Christy Ballweber, Carol Dweck and Zoran Popović. 2014. Brain points: A growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3339-3348. http://dx.doi.org/10.1145/2556288.2557157
42. Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. In *International Conference on Educational Data Mining (EDM 2013)*.
43. Robert Pinsky. 2014. ARPO222x: The Art of Poetry. BUx/Edx. (Sept-Nov 2014). Retrieved from https://www.edx.org/course/art-poetry-bux-arpo222x
44. Jonathan A. Plucker, James C. Kaufman, Jason S. Temple, and Meihua Qian. 2009. Do experts and novices evaluate movies the same way? *Psych & Mark*. 26, 5, 470-478.
45. Michael Prince. 2004. Does active learning work? A review of the research. *Jour of Eng Educ*. 93, 223-232.
46. Chris Quintana, Brian J. Reiser, Elizabeth A. Davis, Joseph Krajcik, Eric Fretz, Ravit Golan Duncan, Eleni Kyza, Daniel Edelson and Elliot Soloway. 2004. A scaffolding design framework for software to support science inquiry. *The Jour of the Learn Sci*. 13, 3, 337-386.
47. Sara Schroter, Nick Black, Stephen Evans, Fiona Godlee, Lyda Osorio and Richard Smith. 2008. What errors do peer reviewers detect, and does training improve their ability to detect them? *Jour of the Royal Soc of Med*. 101, 10, 507-514.
48. Dale H. Schunk and Carl W. Swartz. 1993. Goals and progress feedback: Effects on self-efficacy and writing achievement. *Cont Ed Psych*. 18, 3, 337-354.
49. Keith J. Topping. 2010. Methodological quandaries in studying process and outcomes in peer assessment. *Learn and Instruc*. 20, 4, 339–343. http://dx.doi.org/10.1016/j.learninstruc.2009.08.003
50. J. Gregory Trafton, Erik M. Altmann, Derek P. Brock and Farilee E. Mintz. 2003. Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *Inter Jour of Hu-Comp Stu*. 58, 5, 583-603.
51. Chih-Hsiung Tu and Marina McIsaac. 2002. The relationship of social presence and interaction in online classes. *The Am Jour of Dis Educ*. 16, 3, 131-150.
52. Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Jour of Risk and Uncer*. 5, 4, 297-323.
53. Anne Venables and Raymond Summit. 2003. Enhancing scientific essay writing using peer assessment. *Inno in Educ and Teach Inter*. 40, 3, 281-290.
54. University of California, Berkeley Career Center. Graduate School – Statement [website]. Retrieved November 9, 2014 from https://career.berkeley.edu/Grad/GradStatement
55. Miriam Walker, Leila Takayama and James A. Landay. 2002. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46, 5 (September 2002), 661-665. SAGE Publications.
56. Mark E. Walvoord, Mariëlle H. Hoefnagels, Douglas D. Gaffin, Matthew M. Chumchal and David A. Long. 2008. An analysis of calibrated peer review (CPR) in a science lecture classroom. *Jour of Coll Sci Teach*. 37, 4, 66-73.

57. Xu A, Huang SW, Bailey B. 2014. Voyant: Generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW '14).* ACM, New York, NY, USA, 1433-1444. http://dx.doi.org/10.1145/2531602.2531604

58. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Vennix, Steven P. Dow, Björn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. To appear in *Proceedings of CSCW 2016*.

59. Vivian Zamel. 1985. Responding to student writing. *Tesol Quarterly*. 19, 1, 79-101.