Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection

Julia Cambre[1,2], **Scott Klemmer**[1], **Chinmay Kulkarni**[2]
[1]UC San Diego, [2]Carnegie Mellon University
{jcambre, chinmayk}@cs.cmu.edu, srk@ucsd.edu

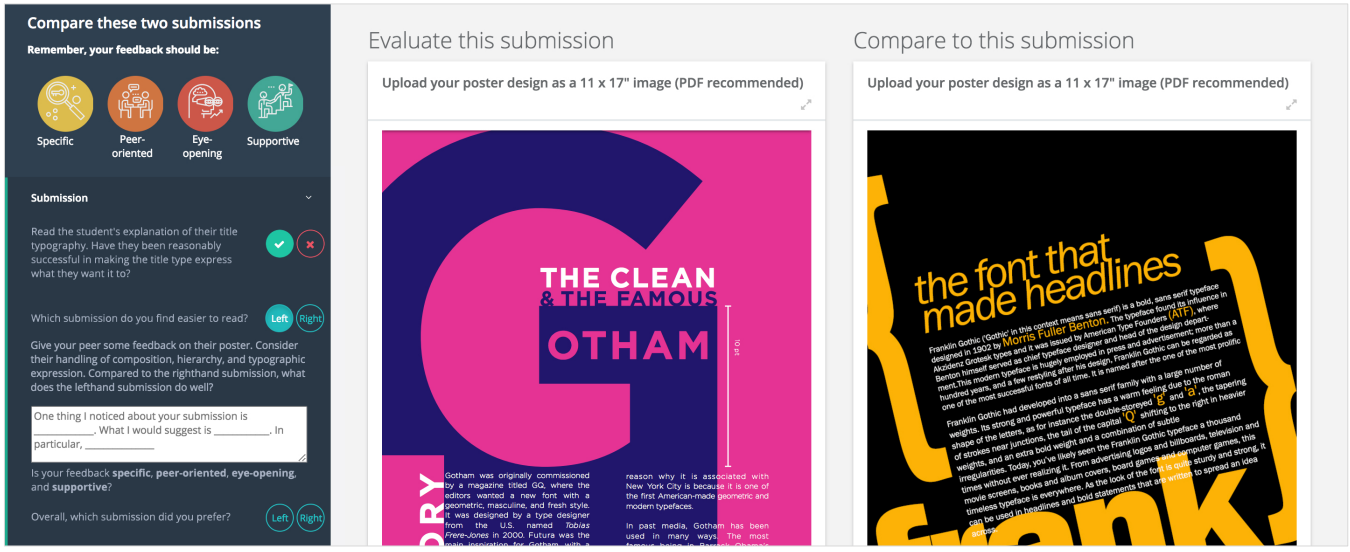

Figure 1. The review interface presents two peer submissions side-by-side, and aligns corresponding submission parts so reviewers can identify similarities and differences. To scaffold feedback, a rubric guides the reviewer through a structured comparison.

ABSTRACT

Peer review asks novices to take on an evaluator's role, yet novices often lack the perspective to accurately assess the quality of others' work. To help learners give feedback on their peers' work through an expert lens, we present the Juxtapeer peer review system for structured comparisons. We build on theories of learning through contrasting cases, and contribute the first empirical evaluation of comparative peer review. In a controlled experiment, 476 consenting learners across four courses submitted 1,297 submissions, 4,102 reviews, and 846 self assessments. Learners assigned to compare submissions wrote reviews and self-reflections that were longer and received higher ratings from experts than those who evaluated submissions one at a time. A second study found that a ranking of submissions derived from learners' comparisons correlates well with staff ranking. These results demonstrate that comparing algorithmically-curated pairs of submissions helps learners write better feedback.

Author Keywords

Comparative peer review; contrasting cases; feedback

ACM Classification Keywords

K.3.1 Computer Uses in Education: Collaborative learning

INTRODUCTION

Both on-campus and online, peer review has reached an unprecedented scale. For example, learners on Coursera create approximately 4,000 new submissions *every day* across the 1,385 courses that use peer review, and most receive a grade within 3 days [Coursera Inc., personal communication]. At its best, peer review provides substantial benefit to both learners and instructors: reviewers gain experience providing constructive feedback on open-ended work [6,32], reviewees receive comments based on the diverse perspectives of multiple peers [10,48], and instructors reduce their feedback and grading burden despite teaching larger classes [20,26,27].

Despite these benefits, peer review still faces a fundamental challenge: learners often do not give expert-quality feedback on their peers' work [25,27,53]. Ineffective feedback has substantial negative consequences; learners are less likely to revise their work [46], and may lower standards [52] or develop a false sense of confidence that degrades future performance because of undeserved praise [22]. There have been many attempts to improve the quality of novice feedback and make it expert-like. So far, the most promis-

ing approaches are templating common feedback for reuse [27], providing interactive hints [24,25], carefully designing rubrics [19,53], and compensating for numerical grading biases using machine learning [26].

However, these approaches fail to address an underlying problem: novices fail to give high quality feedback because they cannot differentiate good work from bad as well as instructors can, or even identify which aspects matter [41]. The peer submissions learners evaluate may be the only examples of an assignment they see apart from their own. By contrast, instructors evaluate student work with significantly more preparation: they have amassed substantial domain knowledge over years of study, and grade many more submissions in a course to give them a holistic sense of the range in quality. More generally, experts organize information more effectively than novices do [18,29], highlight deeper features [9], and infer and articulate more nuanced abstractions [17]. Where novices are misled by salient but superficial features, experts notice deeper structural features [36]. Given these shortcomings, it is hardly surprising that peer feedback lacks the quality of expert critique. We suggest that to overcome these shortcomings, feedback systems should focus not on mechanical aids such as hints or feedback templates, but on helping peers notice features that experts do, infer and articulate expert-like abstractions, and consequently, give better feedback.

This research leverages *structured comparisons* between superficially similar peer submissions as a scaffold for high-quality feedback. Comparisons help novices notice what only experts might otherwise see [44]. For example, when wines are tasted as a flight, or x-rays viewed side-by-side, people can factor out the commonalities and appreciate subtle differences that would go unrecognized if studied individually [15,43].

We hypothesize that comparing peer work will induce more expert-like feedback by guiding learners to develop a more nuanced understanding of peer submissions. To motivate this hypothesis, we draw from theories of learning using a form of comparison known as *contrasting cases*, combine them with best practices from prior research on peer assessment, and empirically test the ensuing concept of *comparative peer review* through a novel software platform, Juxtapeer (Figure 1).

This paper investigates whether *comparative peer review* improves the quality of peer and self assessments. In a randomized controlled experiment with 476 participants across three online courses and one on-campus course, learners assigned to review by comparing two peer submissions wrote feedback that was longer and rated as more specific, deeper, and more likely to use expert terminology than learners who reviewed one submission at a time. Immediately following peer review, learners completed a self reflection using identical interfaces in the two conditions. Those who compared peers wrote reflections that were longer and rated as more specific than the reflections of learners in the serial condition.

In a second study, we evaluate the ranking accuracy of comparative peer review. Using the pairwise preferences that learners submit on each review, an active learning algorithm (Crowd-BT) was able to generate a ranking of submissions by overall quality that correlates well with course staff. Finally, we share learners' qualitative experiences through survey data, and discuss insights to inform future deployments of comparative peer review.

To date, Juxtapeer has enabled comparative peer review in seven courses offered by four different institutions on topics ranging from digital music production to teacher training, drawing a combined learner population from 68 countries (40% United States).

Contrasting cases help develop nuanced understanding

Across domains as diverse as algebra [39], design [12,49], and management [47], learners benefit from articulating the similarities and differences between examples [1]. Juxtaposing two similar examples and studying them closely as *contrasting cases* makes complex structures and subtle features more salient [44]. This also makes learning more efficient by requiring fewer examples to form a schema [21]. Helping learners form a more sophisticated mental model accelerates expertise development [13], which we hypothesize will in turn reduce the gap between novice and expert feedback. Peer review is well-suited to leverage contrasting cases as it is rich in examples. However, incorporating contrasting cases in peer review also introduces three key design challenges.

First, spatial juxtaposition is a prerequisite for effective comparison. Even when similar examples are presented in close succession, learners rarely draw spontaneous connections between them [13,31]; learners perform significantly better when the examples can be compared at a glance [39,44]. In contrast to traditional peer review systems, which ask learners to evaluate each submission on a separate page, Juxtapeer facilitates comparison by juxtaposing two peer submissions side-by-side.

Second, several studies on case-based reasoning have found that even when two cases are presented on the same page, learners often do not identify structural similarities between them unless they are explicitly asked to compare the two [47]. More effortful forms of comparison (e.g. requiring learners to list specific similarities and differences between different cases) also deepen learners' understanding [28]. In state-of-the-art peer review systems, learners rely upon rubrics to guide their feedback [6]. Juxtapeer introduces comparative rubrics. Each item on the rubric prompts learners to choose which submission they think is better along a particular dimension, and to provide feedback based on their comparison.

Third, contrasting cases work best when the juxtaposed cases are maximally similar [14,44]. However, manually

choosing pairs is both tedious and error-prone at large scale. Therefore, Juxtapeer relies upon a machine learning algorithm to choose pairs for review. To find similar submission pairs, the algorithm leverages the comparative judgements on overall submission quality from each review.

Related work in peer review

Juxtapeer builds on a long history of peer review research. Here, we focus on systems and studies with direct relevance to comparative review.

Several studies have leveraged example submissions to improve feedback. For instance, PeerStudio simultaneously presents an example of excellent work in the review interface [25]. While this bears a resemblance to Juxtapeer's design, it is limited in two critical ways. First, selecting an exemplary submission requires substantial effort from course staff; in many cases, there may be multiple submissions that received a perfect score. This design is also poorly suited for new courses or updated assignments, which have no existing pool of submissions to draw from. Second, an exemplary submission may not be the ideal point of comparison, particularly if there is a considerable gap in submission quality (e.g. an A+ submission vs. a C). Similarly, Calibrated Peer Review (C.P.R.) asks learners to grade an example submission at the beginning of the review process and provides feedback on their accuracy relative to an instructor [7,27]. In addition to facing the same curation challenges as PeerStudio, C.P.R. also adds a substantial time burden to reviewers' workload. By contrast, Juxtapeer aims to train learners "on the job" through structured comparisons instead of a separate upfront training step.

Prior work has also investigated relative grading schemes for peer review. One approach is ordinal grading, which asks learners to rank a set of submissions from best to worst. However, prior work on ordinal grading centers around the theoretical viability of accurately ranking submissions (e.g. [23,38,45,51]); Juxtapeer instead emphasizes comparison as a scaffold to help reviewers provide better feedback. Second, the ComPAIR system shows two peer submissions side-by-side. It asks reviewers to provide feedback on each submission individually, then asks students to choose the "better" of the pair [37]. This prior work offers a system prototype and self-report data from users, but no prior work offers an experiment evaluating the efficacy of comparative review. In addition to offering the first such evaluation of comparative peer review, Juxtapeer differs from ComPAIR in anchoring the comparison on one submission at a time, and asking reviewers to evaluate the submission quantitatively as well as qualitatively.

JUXTAPEER SYSTEM DESIGN

We introduce Juxtapeer, a Rails-based web application for comparative peer review in on-campus and online courses.

Juxtapeer organizes peer review around "projects" (assignments), each of which has three phases: submission, review, and results. For continuity, the submission, review, and results interfaces share a common design: a sidebar displays instructions and the evaluation rubric alongside submission material in the main panel (*e.g.,* Figures 1 & 2). Instructors create the instructions, submission structure, and rubrics, and choose the number of peer reviews they would like each learner to perform (typically three or five).

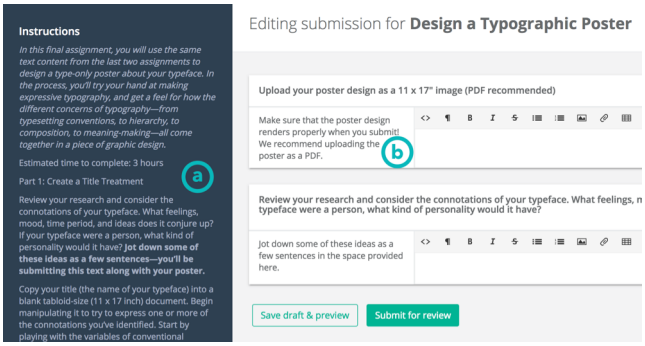

Figure 1. Submission interface in Juxtapeer.

Structuring assignments for alignment

Informed by prior work, Juxtapeer visually aligns examples to be compared, so that perceptual differences are salient, and comparisons are most effective.

Submission: Discrete submission parts standardize structure

To visually align submissions during the review process, they must share a common structure. We designed the submission interface to provide this structure with templates for each submission part. Learners begin by reading project-level instructions and a preview of the review rubric (Figure 1a). Then, they upload their work in several discrete parts, using a dedicated WYSIWYG editor for each (Figure 1b). To encourage revision and minimize errors, learners can save and preview their submission at any time.

Review: Aligned comparisons

Since all submissions share a common structure, Juxtapeer aligns corresponding submission parts in two equally sized columns for review. The learner evaluates only the left-hand submission, using the right-hand submission as a point of comparison. To ensure that submission parts can be easily compared without navigating between tabs or referring to a download, Juxtapeer embeds rich content inline.

On each review, learners complete an evaluation rubric of graded criteria, comparative preferences, and open-ended comments. Juxtapeer allows "Yes/No" binary criteria and open-ended comment fields common in many peer review platforms, and the "comparison" rubric item type. Comparison rubric items prompt learners to examine the submissions along a particular dimension, and to select the submission that best matches the criteria through buttons labeled "Left" and "Right." Once reviewers indicate their preference, the rubric item expands to reveal a comment box with an adaptive prompt.

Receiving feedback: Comparative reviews shared in context

After reviewing peers, learners receive feedback on their own submission. Rather than aggregating feedback from

Figure 3. Comparison learners see two submissions at a time, but evaluate only the submission on the left. In both conditions, students see the same number of submissions, and perform the same number of reviews.

multiple peers, the interface for viewing feedback exactly mirrors the review experience. Submitters browse comparative feedback from each reviewer in context, and see the comparison submission (which may differ between reviews) directly alongside their own. This trades off ease of noticing where reviewers agree for seeing feedback in context. This allows learners to see additional examples of peer work, and make sense of feedback that references a feature of the comparison submission.

Optimal submission pairs for review with Crowd-BT

To pick similar submissions for comparison, Juxtapeer uses a modified version of the Crowd-BT [8] active learning algorithm. Crowd-BT is based on the Bradley-Terry model, which suggests that learners are more likely to make a comparison 'correctly' (i.e. agree with the consensus) if the difference between items is large. Conversely, it can compute the 'reliability' of a rater based on how small a difference they can compare correctly. Over repeated comparisons, Crowd-BT improves estimates of quality and reliability. By weighting more reliable raters higher, and asking raters to compare pairs with smaller perceived differences, Crowd-BT improves ranking accuracy with a given number of comparisons. The final item on each rubric asks learners which submission they prefer overall; Crowd-BT uses this pairwise preference to update its quality estimate of both submissions. Juxtapeer modifies Crowd-BT as implemented by Athalye [3]. To ensure feedback is distributed equitably, we prioritize submissions that have not yet received a target number of reviews, and restrict the pool to submissions that a learner has not yet reviewed.

Giving comparative feedback on a pair of submissions requires students to read and understand both submissions. To reduce cognitive load, Juxtapeer anchors evaluation on the left-hand submission, and exposes learners to only one unfamiliar submission on each new review. As learners move from one review to the next, the submission they just evaluated shifts to the right to become the comparison, and a new submission appears for review on the left (Figure 3).

Scaffolds to encourage quality feedback

In an early pilot of Juxtapeer, we noticed that when asked to write comments after making a comparison, learners often wrote rationale that defended their choice, rather than constructive peer feedback. For example, reviewers wrote about the relative strengths of one submission ("less interesting" or "better"), or pointed out something missing or problematic about the submission they did not select. To encourage more constructive feedback, Juxtapeer scaffolds comments in two ways.

Walkthrough guide on providing high-quality feedback

One common approach to scaffold comments is to provide context-relevant examples (or "fortune cookies") of common feedback [16,24,27,33,53]. Similarly, we show learners a guide for effective comparative feedback before starting their first review on each project. Just as "fortune cookies" leverage the idea that recognition is more reliable than recall [2], Juxtapeer's guide helps students recognize four patterns for good feedback; feedback should be *specific* (comment on details rather than surface-level features), *peer-oriented* (provide constructive feedback to the peer, not justification for the course staff), *eye-opening* (suggest a possible new direction), and *supportive* (include encouragement). A brief panel for each principle outlines simple "do's and don't's", and includes an example of good feedback in that category. If a learner attempts to submit a review containing a comment under 20 characters, we redisplay the guide to remind learners of these principles, and ask them to elaborate on their feedback.

Placeholder text provides sentence starters

How feedback is framed affects how it is received [19,35], and novice feedback providers benefit from guidance on how to structure their comments [24]. Each comment field offers three short sentence starters as placeholders, inspired by the adaptive comment prompts in PeerStudio [25].

STUDY 1: COMPARATIVE PEER REVIEW IN 4 COURSES

To test the efficacy of comparative peer feedback, we conducted a controlled experiment in three massive open online courses and one in-person course. This evaluation sought to address three research questions:

1. Does peer feedback through comparisons yield higher quality feedback than peer feedback provided on one submission at a time?
2. Does comparative assessment affect learners' willingness to engage with peer review or the class?
3. How does comparative peer review affect learners' perceptions of their own work in self assessment?

Methods

Juxtapeer served as the assessment platform for open-ended work across 13 assignments in 4 courses (Table 1). Across courses, 476 students consented to participating in our study; they came from 68 different countries, with 39.9% from the United States.

Study Design

A between-subjects manipulation randomly assigned consenting learners to complete peer reviews on Juxtapeer in one of two conditions: *compare* (manipulation) or *serial* (control). Consenting learners were split evenly between conditions upon submitting their first project, and maintained the same condition throughout the course. In the compare condition, learners reviewed by comparing two

Course	Project Descriptions
Introduction to Typography CalArts Online (Coursera)	Learners began by researching and writing a report on a typeface of their choice. In assignments 2 and 3, they stylized their research using typographic forms, and later converted that composition into a poster design.
Jazz Improvisation Berklee School of Music Online (Coursera)	Learners recorded jazz improvisations on an instrument of their choice, and uploaded their work as an MP3 file. The first 4 projects were optional, and the final project was required.
Pro Tools Basics Berklee School of Music Online (Coursera)	In two projects, learners created short music productions and uploaded screenshots of their mix & edit setup.
Reflective Teaching Practice UC San Diego On-campus (graduate seminar)	Across 3 milestones, learners created and iterated upon lesson plans for their subject area of expertise in elementary and secondary classrooms.

Table 1. Overview of courses in Study 1.

submissions, as described above. In the serial condition, the peer review interface displayed only one submission at a time. The review process was double blind in both conditions; learners did not know the identities of the peers they reviewed, nor of the peers who reviewed them.

In the compare condition, the last rubric item asked learners for an overall comparison and qualitative feedback on the submission as a whole. Learners first indicated which submission they preferred before entering comments. The reviewer's selection to a comparison prompt served as the input for the ranking and review distribution algorithm (Crowd-BT), and also determined the feedback prompt. If the reviewer indicated that they preferred the submission on the left, the prompt asked "In comparison to the submission on the right, what does the submission on the left do well?". If they selected the right submission, the prompt asked how the left submission could be improved.

In the serial condition, the final rubric item asked learners to provide qualitative feedback on the submission as a whole. To keep comment prompts maximally similar across conditions, the serial condition kept a running score of the reviewer's responses to the quantitative rubric items that preceded a comment. If the score for the preceding rubric items was 80% or above, reviewers saw the prompt asking what the submission did well. Otherwise, they were asked how the submission could be improved. The serial condition also used Crowd-BT to select which submissions a learner reviewed. We treated sequential submissions as the "pairs" in the serial condition, and used quantitative scores as a proxy for preference. If the reviewer awarded a higher overall score to the current submission than to the previous submission, the current submission was considered as "preferred"; if the two submissions received the same score, then the preferred submission was chosen at random.

To ensure that both conditions saw the same number of submissions during review, the compare condition kept one submission consistent between reviews (Figure 3). After finishing peer reviews, learners in both conditions used an identical interface to reflect on their own work. The self-assessment interface is similar to the reviewing interface in the serial condition, and displays only the learner's own submission. Comment prompts were the same across both conditions, and used the favorability score as in the serial condition. We also showed identical reviewing guides in both conditions.

After receiving feedback on their first and last projects in a course, consenting learners were invited to complete a brief survey about their experience with Juxtapeer. Survey participation was optional, and did not affect grades.

Analysis of textual data

We coded a random sample of reviews along five dimensions to measure feedback quality in a two-phase process.

To validate the coding scheme and establish inter-rater reliability, the initial phase focused on reviews from the largest course in the dataset, Typography. We randomly selected 40 learners (20 per condition) who completed all three required projects in the course and sampled one of their peer reviews on each project at random. This yielded 120 peer reviews, split evenly between the conditions and projects.

Three human-computer interaction graduate students outside the research team independently coded these 120 reviews, blind to condition. Students had prior knowledge of typography through coursework. On each review, coders indicated whether the feedback was *specific*, *supportive*, *suggested action*, and *used expert terminology*. These dimensions were binary and non-exclusive. Additionally, coders rated the *depth of reflection* on each review on a 7-point scale from "extremely superficial" to "extremely deep." We provided coders with a definition and example for each criteria.

Fleiss' Kappa indicated substantial agreement on the specific dimension (0.64), and near-perfect agreement on the suggest-action (0.86) and supportive (0.92) dimensions. Krippendorff's alpha for review depth ratings was 0.82, indicating substantial agreement. The expert terminology dimension reached only moderate agreement (Fleiss' $\kappa = 0.54$), as raters had different standards for whether certain commonly occurring terms (*e.g.*, "typeface") constituted technical terminology. The same three coders also manually rated self assessment quality. This sample comprised 117 self assessments: one per project for each of the 40 learners whose feedback was coded for peer review, with the exception of three learners who did not complete their self assessment on the third project. Coders were again blind to condition, and indicated whether the self assessments were *self-critical* (Fleiss' $\kappa = 0.91$) and *made reference to other submissions* (Fleiss' $\kappa = 0.95$). Like on the peer assessments, we also asked coders to provide a rating for *depth of reflec-*

tion on a 7-point scale (Krippendorff's α=0.82), and to indicate whether self assessments were *specific*. However, because self assessments are typically written only for the learner's own purposes, and not for an external audience, agreement was moderate for this measure (Fleiss' κ=0.50). For both peer and self assessments, we measured review quality using the median of the three coders' scores on each dimension.

The second phase applied this coding scheme to a wider sample of reviews. A member of the research team independently rated the same 120 peer and 117 self-assessment samples from Typography, and matched the median score derived from the first coding phase with moderate agreement or better on all measures. The researcher then coded reviews from the remaining three courses, blind to condition. To compile these additional review sets, we chose two peer reviews and one self assessment at random from each consenting learner who completed the review requirement for at least one project. Including the Typography reviews, this yielded a final data set of 494 peer reviews (12.0% of total) and 304 self assessments (35.9% of total).

Analysis of quantitative data

For dependent variables such as grades, comment length, and time to complete reviews, we built linear mixed-effects models with a fixed-intercept random effect for course, user identity, and assignment. Comment length and time variables followed a log-normal distribution, and were log-transformed prior to analysis.

Results

We analyzed work submitted by the 476 consenting participants that was required (to avoid selection bias), non-blank (to exclude content submitted in error), and in English (the language of instruction in all classes). This comprised 1,297 submissions, 4,102 reviews, and 846 self assessments.

To account for multiple hypothesis testing and provide a conservative measure of significance, we apply a Bonferroni family-wide error correction based on three families of data: textual coding, quantitative peer-review data, and quantitative self-assessment data. We present the exact p-values after correction. We first present aggregated data from across all courses, and then discuss notable course-specific differences that emerged through analysis. These supplementary individual course analyses are meant to be exploratory, and do not use Bonferroni corrections.

Comparing submissions improves qualitative peer feedback

Learners in the comparative feedback condition provided feedback that was rated as significantly more *specific* ($\chi^2(1)=9.32$, p=0.011). These reviews also outperformed reviews from the serial condition in their *use of expert terminology* ($\chi^2(1)=13.3$, p=0.0013) and *depth of reflection* (Z=3.10, p=0.010). There were no statistically significant differences between conditions in whether the reviews *suggest action* ($\chi^2(1)=3.45$, p=0.32). Though feedback was highly positive in both conditions, comparison-based reviews trended towards being less likely to include *supportive* language ($\chi^2(1) = 4.20$, p=0.20). Figure 4 details the Bonferroni-corrected course-level significance on each dimension.

Comparing peer work also yielded significantly longer feedback: $t(596) = 3.34$, p=0.0036 (Figure 5). The median comment length for peer reviews in the compare condition

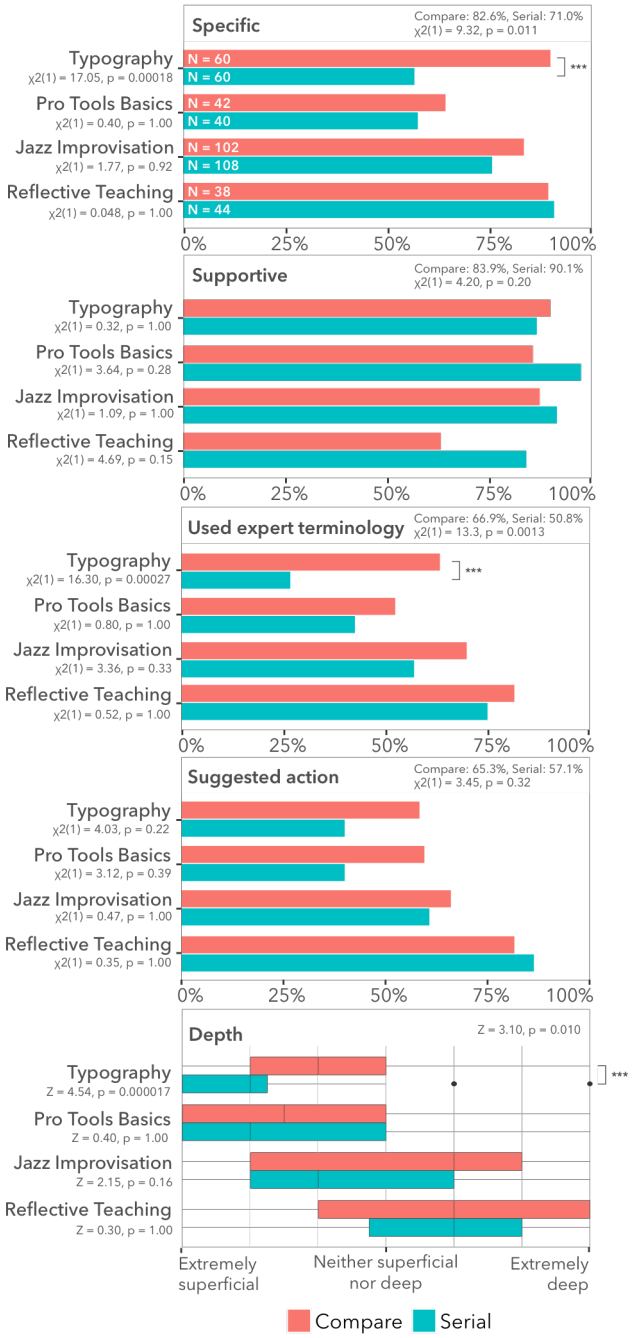

Figure 4. Manual coding of a random sample of peer reviews suggests that comparative reviews outperform serial reviews along all dimensions except supportive language. Depth was coded on a 7-point Likert scale, and the remaining measures were binary and non-exclusive. Aggregate values at the top right; N is number of reviews in each bar.

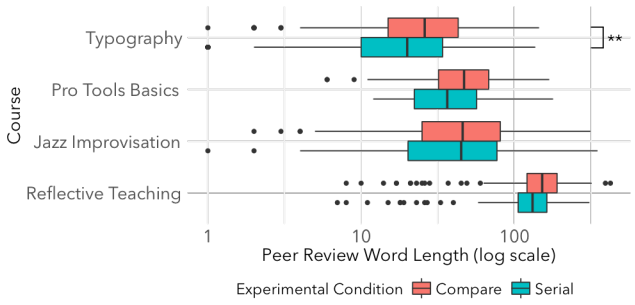

Figure 5. Log-transformed peer review word length by course and condition. Reviews containing more than one comment field were added together to measure the total word length by review. Peer reviews in the compare condition are significantly longer than in the serial condition (t(596)=3.34, p=0.0036).

was 36 words (IQR=50), compared to a median of 27 words (IQR=45) in the serial condition.

Accuracy and consistency of grades

Reviewers in the compare condition assigned scores that were lower than those in the serial condition (t(403)=−2.03, p=0.17), though this difference was not statistically significant. At the individual review level, scores from the compare condition were 1.2% lower than those of the serial condition. Comparison learners were significantly less likely to give a submission full points ($\chi^2(1)=7.68$, $p=0.022$). However, comparison had no impact on the aggregate submission score computed by the median of peer reviews on each submission ($t(944)=-1.48$, $p=0.14$). Overall grades skewed high in both conditions; 77.4% of submissions received a perfect score.

Comparison learners spent more time reviewing

Learners in the compare condition spent significantly longer (t(489)=3.37, p=0.0033) on each review than those in the serial condition (Figure 6). In the three online courses, the median time per review was 4 minutes 2 seconds in the compare condition, and 3 minutes 1 second in the serial condition. Reviews in the in-person Reflective Teaching course took substantially more time (11:20 in compare and 11:05 in serial), likely because the assignments involved longer rubrics and text-based content. For the review completion time analysis, we exclude 425 peer reviews (10.4% of the required reviews) with a total completion duration of over one hour. Such large review completion times were usually the result of learners beginning a review, and returning days later to complete it.

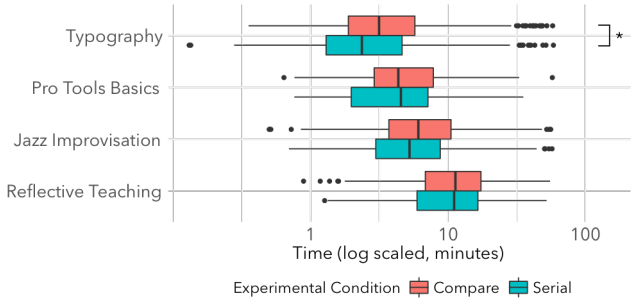

Figure 6. Comparison significantly increased review time.

Learners who compare peer work write longer, higher quality self-reflections

Though the self-assessment interface was identical for both conditions, learners who reviewed peers comparatively went on to write significantly longer self assessments ($t(446)=2.42$, $p=0.032$). The median self assessment in the compare condition was 28 words (IQR=37), and the serial condition wrote a median of 21 words (IQR=34). Reflections from learners in the compare condition were significantly more *specific* ($\chi^2(1)=7.77$, $p=0.021$). While not statistically significant, self assessments from comparison learners tended to be more likely to *refer to other submissions* that the learner reviewed ($\chi^2(1)=4.67$, $p=0.12$) than those in the serial condition, and tended to receive higher ratings on their *depth of reflection* ($Z=2.09$, $p=0.15$). Learners were equally likely to include *self-critical* language ($\chi^2(1)=0.0015$, $p=1.00$). There was no significant difference between the self-assessed scores by condition ($t(321)=-0.076$, $p=1.00$).

Comparison does not significantly impact learner engagement or perceived review quality

Comparing peer work had no significant effect on persistence in the Typography course. We did not analyze persistence in the other online courses because they each involved only one required project. The continuation rate (percentage of learners who submitted two consecutive projects) was 60.2% in the compare condition and 63.7% in the serial condition between the first and second project ($\chi^2(1)=0.33$, $p=0.57$). Of those who submitted to the second project in the compare condition, 58.1% also submitted to the third project, and 67.4% of learners did so in the serial condition ($\chi^2(1)=1.50$, $p=0.22$). Willingness to complete extra reviews (voluntarily) also did not differ significantly ($\chi^2(1)=3.11$, $p=0.08$) between the conditions. Overall, 12.4% of students completed more reviews than required.

When learners viewed their feedback on each review, Juxtapeer solicited input on perceived feedback quality by asking, "Would you like to receive comments similar to the ones in this review on future assignments?" (Yes / No / I don't want to say). Both conditions were equally likely to favor the reviews they received ($\chi^2(1)=1.11$, $p=0.29$); learners approved of the comments on 80.7% of reviews that received a rating (23.1% of all reviews).

Visual work especially benefits from comparison

We found that the effect sizes in the Typography course were much larger than the other three courses. To understand whether Typography had an outsized influence on significance, we re-ran the analyses without reviews from Typography. The statistical results are directionally the same when we remove the course, but some are non-significant due to a smaller dataset. For peer review data, comments were less supportive in the compare condition ($\chi^2(1)=6.94$, $p=0.0084$), used more expert terminology

($\chi^2(1)=3.85$, p=0.050), had marginally greater rated depth (Z=1.62, p=0.11). Similarly, there continues to be no difference in peer feedback that is action-oriented ($\chi^2(1)=1.06$, p=0.30), and self-criticality of self-reflections ($\chi^2(1)=0.0027$, p=0.96). As with the whole dataset, learners also spent marginally more time reviewing in the compare condition (t(263) = 1.84, p=0.067). However, comments were not significantly longer (t(371) = 1.35, p=0.18, and t(207) = 1.06, p=0.29) for peer and self-assessments, respectively. When excluding the Typography course, peer comments were also not significantly more specific ($\chi^2(1)=1.19$, p=0.27). Self-assessments were not significantly more specific ($\chi^2(1)=3.06$, p=0.080), deep (Z=0.74, p=0.46), or likely refer to other submissions ($\chi^2(1)=0.18$, p=0.67). This may suggest that certain content types such as visual work (Typography) are more amenable to comparison than others, such as audio (Jazz Improvisation and Pro Tools Basics) or text (Reflective Teaching).

STUDY 2: CAN COMPARATIVE REVIEW RELIABLY RANK SUBMISSION QUALITY?

Instructors often wish to identify example submissions to recognize outstanding work, or to reuse in a future offering of the course. For example, showing learners an example submission that is slightly better than their own improves future performance more than sharing an outstanding example because learners can envision how to change their work without getting discouraged [25,40]. However, finding the appropriate neighboring submissions would require a complete ranking. Unfortunately, this can be impractical or imprecise at large scale; there may be dozens of submissions which receive full points, so instructors cannot feasibly review and rank all submission themselves.

In the first study, reviewers' pairwise comparisons served to focus the learners' attention on the salient differences between the two submissions. On the back-end, the pairwise judgements also fed into the Crowd-BT algorithm to adjust the rank and confidence of the evaluated submissions. How accurate are the overall pairwise preference judgements in ranking submission quality?

While prior work has looked at using peer assessment to algorithmically rank submissions, it did not actually ask participants to make pairwise comparisons. To our knowledge, we are the first study to evaluate peer ranking through pairwise comparisons with real-world constraints on number of comparisons, human performance, etc. For example, Raman & Joachims [38] use 10-point Likert ratings. With BayesRank [51], learners rank a small set of submissions (e.g. from 1 to 5). Even though these tasks seem similar to comparison, crucial differences may impact the measured accuracy. Rating is done one-at-a-time, and does not benefit from comparison. Ranking may not yield the same results as pairwise comparisons, because other items present distort preferences among two items [50]. Study 2 evaluates how an algorithm for ranking peer reviewed submissions performs in a naturalistic setting.

Instructor Ranking	(Best) 1	2	3	4	5	6	7	(Worst) 8
Crowd-BT	3	2	1	7	6	5	4	8
Teaching Assistant	3	1	2	5	4	7	6	8

Table 2. An algorithmic ranking (Crowd-BT) achieved moderate accuracy against expert rankings in a HCI course.

Instructor Ranking	(Best) 1	2	3	4	(Worst) 5
Crowd-BT	1	2	5	4	3

Table 3. Crowd-BT nearly matched an anthropology instructor's ranking of 5 submissions, with 1 inversion.

Comparative peer review in a group project context

We studied Crowd-BT's ranking accuracy with a modified version of Juxtapeer for group projects in two courses. For one assignment in a large, on-campus introductory human-computer interaction course at a university in Israel, 18 groups of roughly three students each submitted and reviewed storyboards and paper prototypes. As a final project in an on-campus anthropology course offered at an American university, five groups wrote mock grant proposals for HIV prevention and support programs. In both of these courses, one student submitted on behalf of the group; during the evaluation phase, each student was required to complete 3 reviews, followed by a self-evaluation of their group's submission.

After the review period ended, we asked each course's teaching team to independently rank submissions by overall quality, blind to the algorithmic ranking. Submissions were presented in random order and in isolation using the serial interface. In the human-computer interaction course, we selected the best two and worst two submissions according to Crowd-BT's ranking, as well as 4 other submissions at equal intervals (2 positions apart) in between. For anthropology, the instructor ranked all five submissions.

Results

The algorithmic ranking (Crowd-BT) achieved moderate to strong accuracy. With 118 peer reviews on 18 submissions in the human-computer interaction course, the algorithm was able recall the submissions that both the instructor and teaching assistant ranked in the top three positions (corresponding to submissions in the top five overall) (Table 2). Instructor and teaching staff rankings showed very strong agreement (Spearman ρ=0.88), and the algorithmic ranking was strongly correlated with both the instructor (Spearman ρ=0.67) and teaching staff (Spearman ρ=0.79). In the anthropology course, Crowd-BT inverted the submissions in positions 3 and 5, but otherwise matched the instructor's ordering on five submissions with 75 peer reviews (Table 3). The Spearman correlation between the Crowd-BT ranking and the anthropology instructor's ranking was 0.60.

Analysis

While Crowd-BT generally succeeded in clustering submissions in the right vicinity, the accuracy we observed in

these two deployments is likely insufficient for high stakes performance evaluation (e.g. to grade learners on a curve, or to provide recognition for top submissions). Three modifications may yield more useful rankings in the future. One option is to rank submissions not just by their overall quality, but by their merits along specific dimensions. In the present implementation, the "overall" preference may have resulted in noisy inputs to the ranking algorithm. Additionally, tuning the parameter weights may yield a more accurate ranking. For example, the platform could adjust the priors for course staff to reflect higher confidence in their judgements, and ask them to seed the review pool with comparisons before learners begin reviewing. Finally, increasing the number of reviews each submission receives would give Crowd-BT a stronger signal from which to rank.

More generally, algorithms like Crowd-BT introduce powerful new affordances in the peer review process. On Juxtapeer, we chose Crowd-BT primarily for its potential to select and rank submissions. However, just as the algorithm estimates submission quality, it also can estimate reviewers' reliability. This suggests a potential third application of Crowd-BT to improve peer review: if a reviewer's preference judgements seem random or diverge considerably from their peers, the algorithm could flag their behavior for instructors' attention. This metric could also help match reviewers with similar prior experience or knowledge.

DISCUSSION

These studies demonstrate that comparison amplifies the benefits of peer review. Contrasting cases have well-established pedagogical value, yet most prior work involving contrasting cases has used highly curated examples [1,44]. These results suggest that with algorithmically chosen submission pairs and aligned comparisons, contrasting cases can improve peers' feedback, even within a less controlled environment like peer review. Below, we discuss how comparison adds value to the peer review process, and consider the conditions in which it is most successful. Learner responses are based on a survey that consenting participants were invited to complete after receiving feedback on their first and last projects in a course (N=94).

Comparison especially effective for visual submissions

At its best, comparison helped learners notice details of their peer's submission that they might otherwise have overlooked. As one learner wrote: *"I feel it standardizes the evaluation grading. It's great to see 2 submissions side by side, because you have a point of reference to grade. Also, it's a learning stimulation, as you tend to jump quickly from submission A to B and vice versa to identify the good/best answer or to validate if the target submission as the correct answer."* This comment underscores a major strength of contrasting cases: it helps learners recognize criteria for submission quality themselves, rather than requiring instructors to explicitly define them.

However, we also found that effective comparison is contingent upon context. The benefits for comparison were especially pronounced in the Typography course, in which peer reviews were rated as significantly higher in quality on three dimensions (specific, deep, and use of expert terminology). This provides evidence about the boundary conditions for comparative peer review. The visual nature of the Typography projects may be more amenable to comparison. For example, the interface made differences between graphic designs perceptually salient. In contrast, actively comparing two recordings that are several minutes long is harder because the tracks cannot be appreciated simultaneously. Comparative peer review may also have benefits if student work can be aligned symbolically, even if not perceptually. For instance, the music courses in this study may have benefitted by asking learners to also upload an image of their musical score.

In this study, reviewers in the compare condition took significantly longer to complete each review than those in the serial condition. Additionally, learners who compared peer work assigned significantly lower scores, and were less likely to use supportive language. However, the actual differences between conditions were small (median increase of approximately 1 minute in time, 1.2% decrease in scores, and 6.2% difference in supportive language). We speculate that this small additional time per review is the result of students conceptually aligning examples, and may well lead to better learning [13]. Similarly, learners may assign lower scores and write less supportive feedback because comparison makes missing features apparent, thus orienting them as more critical reviewers.

Comparative review generally well-understood, perceived as fair

Comparative peer review is a novel classroom interaction, but was generally well-understood by students. Learners generally rated their experience with Juxtapeer highly regardless of condition; we found no significant differences between conditions on any self-report measures. Of the 1,991 comparative peer reviews, in only three cases did learners report that the feedback they received corresponded to the wrong submission. (One learner reported: "*They reviewed the wrong poster. The comments are for the poster on the right hand side*.") The other failure mode seems to be that some learners assumed both submissions were created by the same peer. For example, one comparative review of different learners' posters read: "*This was a tough one because I love both of your pieces. The only reason I preferred the right was a personal preference of stark contrast. But I absolutely love your use of letters to convey composition and theme*." Future iterations could improve this even further, *e.g.*, by reducing the opacity of the comparison submission to focus attention, or highlighting that submissions were authored by different learners.

In both serial and comparison-based reviewing, learners generally perceived peer reviewing to be fair. On each review, we invited consenting learners to share whether they felt the feedback was helpful, and why. In only 17 of 600

reviews (2.83%) that received a response in the compare condition and 6 of 343 reviews (1.75%) in the serial conditions did learners claim that they were graded inaccurately, pointing to specific rubric items where a reviewer deducted points (differences between conditions are non-significant).

Social comparison in comparative peer review

Comparison is a social process, and has challenging social implications. Even though the peer review process was double blind, we found some evidence that some learners felt uncomfortable making comparisons between classmates' work. For example, learners should be equally like to prefer the left or right submission, but learners were significantly more likely to prefer the submission on the left (i.e. the one their feedback was directed towards), over the comparison submission ($\chi^2(1) = 21.1$, $p < 0.001$; 55.1% of reviews). In line with prior work (*e.g.* [19]), this suggests reviewers are sensitive to how feedback will be perceived.

Similar sentiments regarding comparison emerged in survey responses; two learners remarked that choosing between submissions felt like an unfair or nonsensical task: "*I didn't really like the fact that we need to compare works by two people, it's not always possible to choose. Sometimes the two works are equally good, or bad."* Some students felt that the process failed to allow for such personal preferences: "*The format forces the reviewer to favor one work over the other and as a result the reviewer has to explain what they don't like about the piece they did not choose."* Others were deeply concerned that comparisons were *only* personal preference, not an objective criterion. One wrote: "*Comparative assessment cannot ever lead to fair and accurate assessment of another's work. It is very much like asking someone if they prefer fried or boiled potatoes. Very much a matter of personal taste. Any review process should be aimed at achieving objective results and not results based on likes and dislikes."*

Though we recognize these learner frustrations, we believe that eliciting preferences and asking for rationale create desirable difficulties [4]. To develop professional wisdom, learners need to reflect, and form preferences [42]. Encouraging learners to articulate why they prefer one submission over the other may pave the way to deeper understanding.

Nevertheless, we acknowledge that interaction patterns for comparison are a difficult design challenge. Some educators explicitly advise against any form of comparison in peer feedback, citing potential negative consequences for interpersonal relations and intrinsic motivation [5], and heightened risk of drop out in online courses [40]. However, such social comparisons may also build social proof to encourage positive learning habits [11], and help to match learners with potential mentors [30,34]. In a follow-up survey, 53.6% of learners indicated that they were interested in connecting with the peers with whom they exchanged feedback on Juxtapeer. Future work should explore when social comparison is helpful, and how best to leverage it.

LIMITATIONS

We see two main limitations of the current work. First, we did not measure the qualitative similarity of the submission pairs that Crowd-BT chose; coding the submissions for common features may provide a better understanding of the algorithm's performance, and of how the degree of submission alignment mediates feedback quality. Second, these results do not quantitatively measure learning gains or transfer. Going forward, researchers should collaborate with instructors early in course development to design pre- and post-tests to better quantify the benefits of providing and receiving comparative feedback. Despite these limitations, these results show that comparative peer review is effective in contexts that are diverse in topic, size, and structure.

FUTURE WORK

These results suggest several exciting avenues for future research, both in exploring a broader range of variations on comparative review, and in gaining a deeper understanding of when and how comparison benefits learning. For example, what characteristics of submission pairs generate the highest quality feedback? Juxtapeer pairs submissions by overall quality, yet reviewers may generate even better feedback when they compare submissions that share as many surface-level features as possible. Future studies should investigate whether pairing submissions based on content or quality along more specific dimensions yields better feedback. Additionally, how might we leverage comparison in self assessment? We found that learners who compared peer work wrote longer, more specific reflections on their own work, even without explicitly comparing their work to others. Future work should investigate how comparison could further improve self reflections.

CONCLUSION

This paper demonstrates how theories of learning through contrasting cases can be applied to peer review, and provides empirical evidence that comparison helps reviewers give better feedback on peers' work, and more deeply reflect on their own. To scaffold comparative review at scale, we introduce Juxtapeer, an online platform which has been evaluated in seven courses from four institutions. Learning through examples has traditionally required careful expert curation. These results point to a future where comparing algorithmically curated examples can yield similar benefits in more diverse contexts. For example, job seekers could compare resumes from successful applicants, or newsreaders could study coverage of the same story from multiple sources. In peer review and in the real world, large corpora of examples can help develop nuanced understanding.

ACKNOWLEDGMENTS

We thank Dredge Byung'chu Kang, Sherice Clarke, Jessica Cauchard, Leah Waldo, and the learners who used Juxtapeer for their patience and suggestions, as well as the Coursera team for their support. This work was funded through ONR grant N00014-17-1-2428, and the Carnegie Mellon's Manufacturing Futures Initiatives.

REFERENCES

1. Louis Alfieri, Timothy J. Nokes-Malach, and Christian D. Schunn. 2013. Learning Through Case Comparisons: A Meta-Analytic Review. *Educational Psychologist* 48, 2: 87–113. https://doi.org/10.1080/00461520.2013.775712
2. J.R Anderson and G. H Bower. 1972. Recognition and Retrieval Process in Free recall. *Psychological review* 79, 2: 97–123. https://doi.org/10.1037/h0021465
3. Anish Athalye. 2016. Gavel. Retrieved from https://github.com/anishathalye/gavel
4. Elizabeth L Bjork and Robert A Bjork. 2011. Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*: 55–64. https://doi.org/10.1017/CBO9781107415324.004
5. David Boud. 1995. *Enhancing learning through self-assessment*. Routledge.
6. David Boud, Ruth Cohen, and Jane Sampson. 2014. *Peer learning in higher education: Learning from and with each other*. Routledge.
7. P A Carlson and F C Berry. 2003. Calibrated Peer Review and assessing learning outcomes. *Frontiers in Education Conference* 2: 1–6.
8. Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*: 193–202. https://doi.org/10.1145/2433396.2433420
9. Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. 1981. Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science* 5, 2: 121–152. https://doi.org/10.1207/s15516709cog0502_2
10. Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education* 48, 3: 409–426. https://doi.org/10.1016/j.compedu.2005.02.004
11. Robert B Cialdini. 2001. Influence: Science and Practice. *Book* 3rd: 262. https://doi.org/10.2307/3151490
12. Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction* 17, 4: 1–24. https://doi.org/10.1145/1879831.1879836
13. Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology* 95, 2: 393–408. https://doi.org/10.1037/0022-0663.95.2.393
14. Dedre Gentner and Arthur B Markman. 1994. Structural Alignment in Comparison: No Difference Without Similarity. *Psychological Science* 5, 3: 152–159. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x
15. Eleanor Jack Gibson. 1969. *Principles of perceptual learning and development.* Appleton-Century-Crofts.
16. Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. *C&C 2015*: 1–10. https://doi.org/10.1145/2757226.2757249
17. Brian Greer. 1997. Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction* 7, 4: 293–307. https://doi.org/10.1016/S0959-4752(97)00006-6
18. Adriaan D De Groot, Fernand Gobet, and Riekent W Jongman. 1996. *Perception and memory in chess: Studies in the heuristics of the professional eye.* Van Gorcum & Co.
19. Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*: 458–469. https://doi.org/10.1145/2858036.2858195
20. David A. Joyner. 2017. Scaling Expert Feedback. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale - L@S '17*: 71–80. https://doi.org/10.1145/3051457.3051459
21. Philip J. Kellman, Christine M. Massey, and Ji Y. Son. 2010. Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science* 2, 2: 285–305. https://doi.org/10.1111/j.1756-8765.2009.01053.x
22. Avraham N. Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119, 2: 254–284. https://doi.org/10.1037/0033-2909.119.2.254
23. Pushkar Kolhe, Michael L. Littman, and Charles L. Isbell. 2016. Peer Reviewing Short Answers using

Comparative Judgement. *Proceedings of the Third ACM Conference on Learning @ Scale*: 241–244. https://doi.org/10.1145/2876034.2893424

24. Markus Krause, Tom Garnack, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Critiques using a Natural Language Model. *CHI '17 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May. https://doi.org/10.1145/3025453.3025883

25. Chinmay E. Kulkarni, MS Michael S Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*, 75–84. https://doi.org/10.1145/2724660.2724670

26. Chinmay E. Kulkarni, Richard Socher, Michael S. Bernstein, and Scott R. Klemmer. 2014. Scaling short-answer grading by combining peer assessment with algorithmic scoring. *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*: 99–108. https://doi.org/10.1145/2556325.2566238

27. Chinmay Kulkarni, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and Self Assessment in Massive Online Classes. 20, 6. https://doi.org/10.1007/978-3-319-19641-1

28. Kenneth J. Kurtz, Chun-Hui Miao, and Dedre Gentner. 2001. Learning by Analogical Bootstrapping. *Journal of the Learning Sciences* 10, 4: 417–446. https://doi.org/10.1207/S15327809JLS1004new_2

29. Marianne LaFrance. 1989. The quality of expertise: implications of expert-novice differences for knowledge acquisition. *ACM SIGART Bulletin*, 108: 6–14. https://doi.org/10.1145/63266.63267

30. Penelope Lockwood and Ziva Kunda. 1997. Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology* 73, 1: 91–103. https://doi.org/10.1037/0022-3514.73.1.91

31. Jeffrey Loewenstein, Leigh Thompson, and Dedre Gentner. 1999. Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review* 6, 4: 586–597. https://doi.org/10.3758/BF03212967

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

33. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*: 473–485. https://doi.org/10.1145/2675133.2675283

34. Thomas Mussweiler. 2003. Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review* 110, 3: 472–489. https://doi.org/10.1037/0033-295X.110.3.472

35. Duyen T Nguyen, Thomas R Garncarz, Felicia Ng, Laura Dabbish, and Steven Dow. 2016. Fruitful Feedback: Positive affective language and source anonymity improve critique reception and work outcomes. *Manuscript submitted for publication*: 1024–1034. https://doi.org/10.1145/2998181.2998319

36. Laura R Novick. 1988. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 3: 510.

37. Tiffany Potter, Letitia Englund, James Charbonneau, Mark Thomson MacLean, Jonathan Newell, and Ido Roll. 2017. ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback. *Teaching & Learning Inquiry* 5, 2: 89–113.

38. K Raman and T Joachims. 2014. Methods for ordinal peer grading. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1037–1046. https://doi.org/10.1145/2623330.2623654

39. Bethany Rittle-Johnson and Jon R. Star. 2007. Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology* 99, 3: 561–574. https://doi.org/10.1037/0022-0663.99.3.561

40. T. Rogers and A. Feller. 2016. Discouraged by Peer Excellence: Exposure to Exemplary Peer Performance Causes Quitting. *Psychological Science* 27, 3: 365–374. https://doi.org/10.1177/0956797615623770

41. D.Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Studies* 18, 2: 119–144.

42. Donald A Schön. 1984. *The reflective practitioner: How professionals think in action*. Basic books.

43. Daniel L. Schwartz, Catherine C. Chase, Marily A.

Oppezzo, and Doris B. Chin. 2011. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology* 103, 4: 759–775. https://doi.org/10.1037/a0025140

44. Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. 2016. *The ABCs of How We Learn: 26 Scientifically Proven Approaches, How They Work, and When to Use Them*. WW Norton & Company.

45. Nihar B Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. 2013. A Case for Ordinal Peer-evaluation in MOOCs. *NIPS Workshop on Data Driven Education*: 1–8. Retrieved from http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf

46. Nancy Sommers. 1982. Responding to student writing. *College composition and communication* 33, 2: 148–156.

47. Leigh Thompson, Dedre Gentner, and Jeffrey Loewenstein. 2000. Avoiding Missed Opportunities in Managerial Life: Analogical Training More Powerful Than Individual Case Training. *Organizational Behavior and Human Decision Processes* 82, 1: 60–75. https://doi.org/http://dx.doi.org/10.1006/obhd.2000.2887

48. David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-Based Software Supporting Peer Critique in Large Creative Classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1: 29–35. Retrieved from http://lttf.ieee.org/issues/january2013/Tinapple.pdf

49. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems* 1: 1243–1252. https://doi.org/10.1145/1124772.1124960

50. Amos Tversky and Itamar Simonson. 1993. Context-dependent preferences. *Management science* 39, 10: 1179–1189.

51. A.E.a Waters, D.b Tinapple, and R.G.a Baraniuk. 2015. BayesRank: A bayesian approach to ranked peer grading. *L@S 2015 - 2nd ACM Conference on Learning at Scale*: 177–183. https://doi.org/10.1145/2724660.2724672

52. David Scott Yeager, Valerie Purdie-Vaughns, Julio Garcia, Nancy Apfel, Patti Brzustoski, Allison Master, William T. Hessert, Matthew E. Williams, and Geoffrey L. Cohen. 2014. Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General* 143, 2: 804–824. https://doi.org/10.1037/a0033906

53. A Yuan, K Luther, and M Krause. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. *(CSCW'16) Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*: 1005–1017. https://doi.org/10.1145/2818048.2819953