Interactive Guidance Techniques for Improving Creative Feedback

Tricia J. Ngoon[1], C. Ailie Fraser[1], Ariel S. Weingarten[1], Mira Dontcheva[2], & Scott Klemmer[1]

[1]Design Lab, UC San Diego
La Jolla, CA 92093
{tngoon, cafraser, aweingar, srk}@ucsd.edu

[2]Adobe Research
Seattle, WA 98103
mirad@adobe.com

ABSTRACT

Good feedback is critical to creativity and learning, yet rare. Many people do not know how to actually provide effective feedback. There is increasing demand for quality feedback–and thus feedback givers–in learning and professional settings. This paper contributes empirical evidence that two interactive techniques – reusable suggestions and adaptive guidance–can improve feedback on creative work. We present these techniques embodied in the CritiqueKit system to help reviewers give specific, actionable, and justified feedback. Two real-world deployment studies and two controlled experiments with CritiqueKit found that adaptively-presented suggestions improve the quality of feedback from novice reviewers. Reviewers also reported that suggestions and guidance helped them describe their thoughts and reminded them to provide effective feedback.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

Author Keywords:

feedback; critique; creativity; educational technology

INTRODUCTION: FEEDBACK'S HIDDEN POTENTIAL

Feedback is one of the most powerful influences on learning and achievement [14]. Both giving and receiving formative feedback encourage self-reflection and critical thinking on one's work [24,31], especially in creative and open-ended domains such as design and writing [14,35]. The growing scale of many educational and professional settings increases both the importance and difficulty of providing sufficiently descriptive and personalized feedback. Good feedback can be hard to generate, and people are not consistently skilled in doing so [22,46]. Feedback is often too short, vague, and not actionable [20,40,45]. Even experienced reviewers don't always recognize when they are providing poor feedback or why it is ineffective [40].

CHI 2018, April 21–26, 2018, Montreal, QC, Canada
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5620-6/18/04□$15.00
https://doi.org/10.1145/3173574.3173629

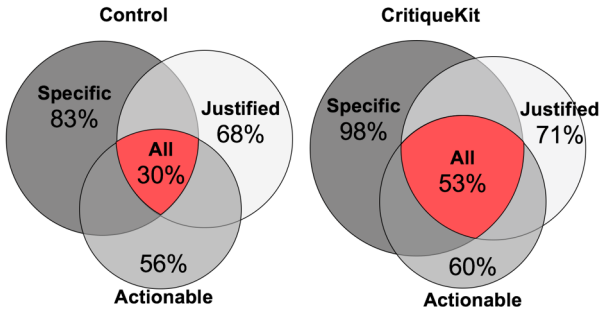

Figure 1. In a controlled experiment, a significantly higher percentage of feedback in the CritiqueKit condition (53% versus 30%) contained three attributes of good feedback: Specific, Actionable, and Justified.

This paper contributes two interactive techniques that improve feedback, their embodiment in the CritiqueKit system, and their evaluation through two deployments and two experiments.

Interactive guidance of feedback characteristics. CritiqueKit features a guidance panel with checkboxes that update as the reviewer gives feedback. A text classifier categorizes feedback as Specific, Actionable, and/or Justified as the reviewer types, providing them with an ambient awareness of their feedback quality and guiding them to improve their feedback.

Suggesting prior feedback for reuse. CritiqueKit enables reviewers to reuse expert feedback, reducing experts' labor by scaling their feedback to similar work. These suggestions update and adapt based on the feedback's categorization to give reviewers targeted ideas for how to improve their comment and provide inspiration.

Two deployment studies and two controlled experiments investigated the efficacy of these interactive techniques on the quality and characteristics of feedback. The first deployment examined how experienced reviewers (teaching assistants) reuse feedback in an undergraduate course. The second deployment examined how undergraduate students reuse feedback. The first experiment examined the impact of statically presented suggestions and interactive guidance on novice feedback. Finally, the second experiment examined the efficacy of adaptively updating suggestions in tandem with interactive guidance on novice feedback. We found that adaptively-presented suggestions improved feedback quality (Figure 1). Reviewers found suggestions useful for inspiration, and the interactive guidance reminded them to ensure

their comments met the criteria for effective feedback. This work provides evidence that interactive techniques such as suggestions and guidance can effectively scaffold the feedback process (See Table 1 for details).

RELATED WORK

Good Feedback is Actionable, yet Rare

Rapid iteration is critical to the success of creative projects, from essays, to visual design, to buildings [5,35]. Receiving feedback early on is important for learners to test alternatives and course-correct [5,41]. Effective feedback is especially important in educational settings where novices are learning new skills and developing expertise. However, giving effective feedback is rarely taught [30]. As physical and digital classrooms increase in size, the demand for feedback outgrows the ability to adopt the ideal learning model of one-to-one feedback [2]. Instead, a one-to-many approach is utilized, where an expert provides feedback for multiple learners. Although learners most value expert feedback [9,27], the one-to-many approach is highly demanding on experts, and specific, actionable feedback for individuals becomes increasingly rare.

In general, effective feedback is specific, actionable, and justified. Specific feedback is direct and related to a particular part of the work rather than vaguely referent [19,35,46]. Specific positive feedback also highlights strengths of the work and provides encouragement, so the recipient can tell they are on a good path [18,43,46]. Actionable feedback is important because it offers the learner a concrete step forward [35,40,43,46]. Simply pointing out a problem is not sufficient to help one improve [32,35,40,41]. Actionable feedback is often most helpful early in a project [4,43,46] because it may help people self-reflect and self-evaluate their work [8], prompting more revisions for improvement [6,42]. Lastly, justification is an important characteristic of feedback [19,28,46], but is arguably one of the hardest to understand or recognize [9]. Justified feedback contains an explanation or reason for a suggested change, which helps the learner understand why the feedback was given.

Study	Adaptive Reuse	Static Reuse	Interactive Guidance	Participants	n=	Main Finding
DEP1		X		TAs	8	TAs used suggestions as inspiration
DEP2		X	X	Design Students	29	Students reused vague suggestions
EXP1		X	X	Design Students	40	Static suggestions and guidance were not helpful
EXP2	X		X	General Students	47	Adaptive suggestions and guidance were helpful

Table 1. Two deployments (DEP) and two between-subjects experiments (EXP) examined the efficacy of feedback reuse and interactive guidance. We found that interactive suggestions and guidance were most helpful for improving feedback.

Rubrics & Examples Usefully Focus Feedback

Rubrics [1,46] and comparative examples [19] are effective in structuring feedback because they beneficially encourage attention to deep and diverse criteria. Novices otherwise tend to focus on the first thing they notice, often surface-level details [12,17,20,46]. Viewing examples of past designs can lead to greater creativity and insights [21,26]; thus, showing examples of good feedback may spark ideas reviewers would not have otherwise considered [12,22,25]. Also, adaptive examples curated to match design features are more helpful than random examples in improving creative work [23].

Rubrics and other scaffolds require significant upfront manual work by experts who must carefully design a comprehensive rubric, curate a thorough set of examples, or decide how else to structure the feedback process. This paper investigates leveraging existing feedback to dynamically create rubric criteria. We hypothesize that showing reviewers previously-provided feedback can guide their attention to important aspects of the design.

Is Feedback too Context-specific for Practical Reuse?

Schön persuasively argues that effective feedback should be context-specific and expert-generated [36]. He offers a vignette from architecture where the teacher suggests an alternative building to the student as an example of situated wisdom and its transfer. If Schön is right that this exchange requires both wisdom and context, does that mean that feedback reuse is infeasible? Within a given setting, project, or genre, common issues recur. Hewing to the principle of recognition over recall, we hypothesize that suggestions and guidance can increase novices' participation in context-specific exchanges.

Prior Systems & Approaches for Scaling Feedback

Existing approaches for scaling personalized feedback include clustering by similarity (*e.g.*, for writing [3] and programming [10,15]). Gradescope [39] and Turnitin [47] allow graders to create reusable rubric items and comments to address common issues and apply them across multiple assignments. Gradescope binds rubric items to scores, which emphasizes grades rather than improvement.

Other methods include automating the reuse of the solutions of previous learners. These methods work best when correct and incorrect solutions are clearly distinct, such as in programming [11,13] and logical deductions [7]. Automated methods have also found success with the formal aspects of more open-ended domains such as writing [3,34]. However, assessing the quality and effectiveness of creative work–the strength of a design, the power of a poem–is intrinsically abstract and subjective and lies beyond current automated analysis techniques. Also, little automated analysis exists for media other than text. For domains like design, human-in-the-loop analysis will remain important for quite some time.

Automatically Detecting Feedback Characteristics

Although feedback is often specific and contextual [36], general characteristics can be automatically detected and used to help reviewers improve their feedback. For example,

PeerStudio detects when comments can be improved based on the length of the comment and the number of relevant words [20]. Data mining and natural-language processing techniques can also automatically detect whether a comment is actionable or not, and prompt the reviewer to include a solution [29,45]. Krause *et al.* use a natural-language processing model to detect linguistic characteristics of feedback and suggest examples to reviewers to help them improve their comment [19]. These methods require a reviewer to first submit their comment so it can be analyzed, and then improve their comment after submission.

CRITIQUEKIT: INTERACTIVELY GUIDING FEEDBACK

Based on these methods and insights, CritiqueKit categorizes feedback and provides prompts and suggestions to reviewers. It differs from prior work by providing feedback to reviewers as they type rather than after they submit. We hypothesize that this ambient feedback with suggestions may provide a just-in-time scaffold that changes how reviewers' thoughts crystallize, yielding feedback that is more specific, actionable, and/or justified.

Interactive Guidance as a Form of Scaffolding

CritiqueKit features an interactive guidance panel with checkboxes that update based on which of three attribute categories the feedback fits: *Is Specific* [19,35,40,46], *Is Actionable* [6,8,20,25,35,43], and *Is Justified* [9,19,28,46].

The prototype assesses the feedback's fit with the following heuristics. The heuristic for the *specific* category merely requires that comments be at least five words long because vague comments tend to be short, such as "good job" or "needs work." Perhaps surprisingly, we observed that the five-word nudge was sufficient to garner specific feedback in practice. (Some websites, like Etsy, also use a five-word minimum heuristic for reviews). For the *actionable* and *justified* categories, we manually combed feedback that had been hand-labeled as meeting these categories and observed that specific keywords (*i.e.,* "maybe try" and "you should" for actionable; "because" and "so that" for justified) were strong cues of these features. Consequently, the prototype implementation simply checks for the presence of these keywords and phrases in feedback comments.

A comment is considered complete once all checkboxes are checked. Reviewers can manually check and uncheck the checkboxes if they feel the checkboxes did or did not add a category in error. For example, if a reviewer's comment states, "Use a 2-column grid layout," and the "Is Actionable" checkbox remains unchecked, the reviewer can manually check the checkbox to note that their comment does indeed contain an actionable suggestion.

Adaptive Suggestions for Greater Specificity

The suggestions box contains a list of previously given feedback from experts. These suggestions dynamically adapt based on how the reviewer's feedback is categorized in the guidance panel. For example, if a reviewer's comment does not yet satisfy the actionable and justified categories (as in Figure 2), the suggestions box would contain examples of feedback with these characteristics. Suggestions appear in the order they were added to the corpus.

The CritiqueKit Review Workflow

When a reviewer first opens CritiqueKit, a prompt asks them to provide specific feedback on something they like about the design and something that could be improved. The suggestions box contains general feedback snippets [22] pertinent to the review criteria to give reviewers a starting point, providing suggestions that are broadly applicable and fit

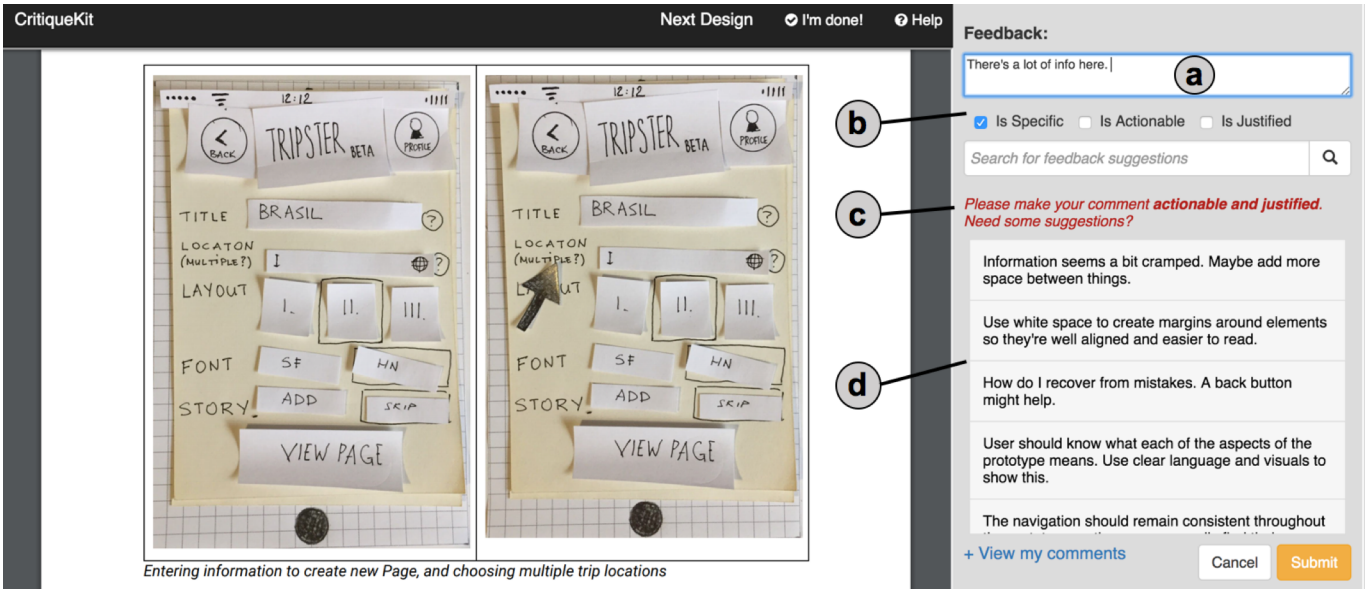

Figure 2. The final CritiqueKit interface for EXP 2. a) The reviewer can type their feedback in the textbox. b) The checkboxes in the guidance panel update based on the characteristics of the reviewer's comments. c) CritiqueKit explicitly prompts reviewers to ensure their comment fits the checkboxes in the guidance panel. d) The reusable feedback suggestions in the suggestions box update based on the unchecked characteristics in the guidance panel, adapting the suggestions specifically to the reviewer's feedback.

within the specified criteria. The "Submit" button at the bottom of the interface is red to indicate that the comment text box is either empty or does not fit any of the categories in the guidance panel.

Once a comment is sufficiently long, the "Is Specific" checkbox will check, and the reviewer will be prompted to make their comment actionable and justified. The "Submit" button turns yellow to indicate that their feedback is not yet complete, though they can still submit if desired. The feedback suggestions then change to present comments that instantiate both actionable and justified feedback. The suggestions continue to adapt depending on the characteristics of the comment, showing reusable examples of feedback that satisfy the unchecked categories in the guidance panel. Once all checkboxes are checked, the "Submit" button turns green as an indication of completeness.

Using prior feedback as suggestions can give inspiration and highlight common issues. The presence of the structured guidance panel reminds reviewers of attributes that feedback should have.

Implementation

CritiqueKit is a client-server web application implemented using Node.js; it assumes that all content to be reviewed is available on the web. The corpus of reusable feedback comments is stored on the server in JSON format.

CritiqueKit uses web sockets for communication between each client running the app and the main server, implemented using the socket.io module. Feedback classification happens on the client-side using JavaScript. Feedback suggestions are also generated on the client-side after retrieving the corpus from the server; the suggestions box adaptively shows and hides comments using JavaScript.

Users access CritiqueKit by navigating to its URL in a web browser. The first time the browser loads the website, a unique ID is generated for the user and sent to the server. A cookie is also saved on the client-side so that the server can identify and differentiate users. The review content is loaded within the page as an iframe.

DEPLOYMENTS: (HOW) IS FEEDBACK REUSED?

To understand how feedback is reused in educational settings and evaluate the CritiqueKit approach, we conducted two deployments and two experiments. All studies took place at a research university.

DEP 1: How Do Teaching Assistants Reuse Feedback?

Eight teaching assistants (TAs) (two female) for an undergraduate design course used Gradescope to grade and critique seven weekly assignments that varied in content from storyboards to written explanations to high-fidelity web application prototypes. TAs set rubric items for each assignment and wrote comments for each. We deployed CritiqueKit to first understand how TAs might reuse feedback and made iterative improvements to the design throughout the quarter based on TA input.

Method: Integrating CritiqueKit with Gradescope

To integrate with the TAs' existing workflow, we implemented CritiqueKit as a Google Chrome extension that augments the Gradescope interface with a suggestions box (Figure 3). This version of CritiqueKit contained only the suggestions box to explore feedback reuse. The suggestions box contained a manually curated set of feedback provided by former TAs in a previous iteration of the course, stored in a Google Sheet online and retrieved by the Chrome extension using the Google Sheets API. Suggestions were categorized into three feedback categories: Positive, Problem, and Solution. TAs could select feedback suggestions to directly copy into the textbox for further editing. Each rubric item contained its own suggestion box interface, providing suggestions specific to that rubric item.

We curated the reusable suggestions corpus as follows. Given all feedback from the previous quarter, feedback that was 25 or fewer words in length was kept, because longer feedback was both too long to be skimmed in a suggestion display and tended to be overly specific. Feedback of 26-30 words was truncated at the sentence level to fit within the 25-word limit. Longer comments or duplicate comments were discarded. In total, 526 comments were provided as suggestions throughout the course for seven (of ten) assignments. Suggestions were manually categorized into the Positive (n=92), Problem (n=312), and Solution (n=122) categories.

Result: TAs Used Feedback Suggestions as Inspiration

Across seven assignments, four of the eight TAs reused 51 distinct suggestions from the 526-element corpus (9.7%). 75

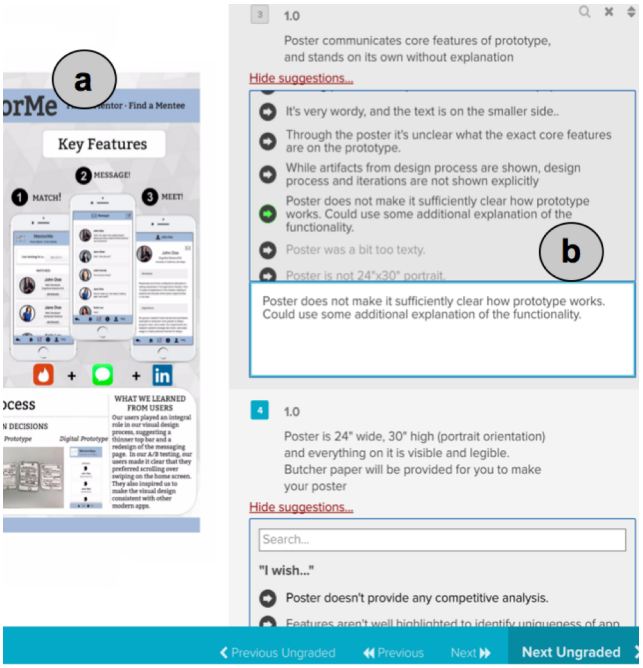

Figure 3. CritiqueKit implemented as a browser extension in Gradescope for DEP 1. a) Reviewers provide feedback on a student design. b) The suggestions box under each rubric item provides reviewers with a list of reusable suggestions and a comment box for providing feedback on a submission.

of 583 designs received a reused suggestion for feedback. 60% of reused suggestions were categorized in the Problem category. These numbers omit any reuse occurring entirely inside Gradescope without CritiqueKit. (Gradescope provides an interface for reusing entire comments within an assignment rather than for individual parts of the comment.)

An end-of-course survey asked TAs about their CritiqueKit use. One commented that he would "*skim the comments in the [suggestions] to see if something was accurate to my thoughts.*" Another mentioned that the prototype helped him *"[find] ways to better explain and give feedback about specific points.*" TAs also mentioned that suggestions sometimes reminded them to comment on more diverse aspects of students' work. For example, one mentioned that seeing positive suggestions reminded her to give positive feedback, not only critiquing areas for improvement. TAs mentioned using the suggestions as inspiration rather than the exact wording, taking the underlying concept of a suggestion and tailoring it.

DEP 2: How Do Students Reuse Feedback?

The first deployment examined teaching staff usage; this second deployment examined student usage to understand how novices interact with guidance and suggestions. We deployed CritiqueKit as a standalone web application with 29 students in an undergraduate design course for five weeks. Students gave anonymous feedback on two randomly assigned peer submissions for each of seven assignments.

Method: Integrating Interactive Guidance for Scaffolding

Novice students are less experienced in giving feedback and may benefit from interactive scaffolding [33]. This version of CritiqueKit included an interactive guidance panel to help reviewers provide more specific and actionable feedback (Figure 4). The categories on the guidance panel were "Is Positive," "Is Specific," "Identifies a Problem," and "Presents a Solution" with checkboxes next to each. These categories stem from recommendations in the literature for both positive and critical feedback [18]. Similar to the final version of CritiqueKit, these checkboxes updated as a reviewer typed by classifying their comment into the three categories. The categories differed from the final version, focusing on specific and actionable feedback.

The suggestions box was seeded with feedback from the course TA. Similar to the first deployment, the suggestions were categorized in the Positive, Problem, and Solution categories. When a student submitted a comment, it was classified into one of these categories, shortened to 25-words if it was longer, and fed back into the corpus to appear as a suggestion, enabling students to reuse their peers' as well as their own comments. The suggestions were ordered first by frequency used, then by shortest length first, and updated as these values changed and more comments were added. Compared to the final version of CritiqueKit, suggestions were static, meaning they did not change as the reviewer typed.

Results: Positive Feedback Common; Reuse Rare

For seven assignments, 898 comments were submitted. Independent raters classified each comment into the five categories of Positive Only, Positive and Specific (Positive + Specific), Problem Only, Solution Only, and Problem with a Solution (Problem+Solution). 45% of these comments contained positive feedback; 30% contained a Problem + Solution statement.

Students rarely selected feedback suggestions for reuse. Over the five-week deployment, 14 distinct suggestions were reused on 27 student designs for four of the seven assignments. These suggestions were mostly short, vague comments such as "I wish this was more visually appealing." This may be

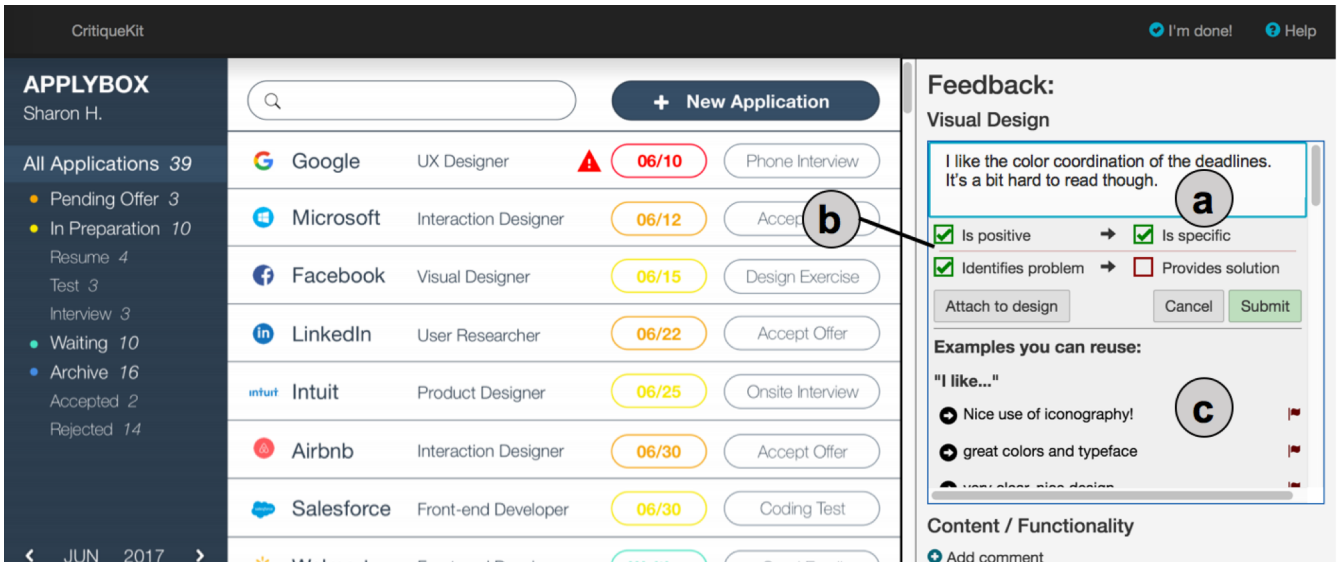

Figure 4. The CritiqueKit user interface for EXP 1. a) The reviewer types their feedback into the text box. b) Checkboxes in the guidance panel update as the reviewer types to show how well the comment fulfills high-quality feedback criteria. c) The reviewer can browse and reuse previously given feedback.

because students often left feedback specific to individual designs that did not easily generalize to other contexts. Students' comments in a post-survey confirmed that the suggestions did not always seem applicable. Students also did not regularly use the interactive guidance panel; 15 of the 29 students engaged with the panel a total of 120 times over five weeks.

In contrast to how TAs reused feedback, students may not have recognized common issues. TAs paid attention to common errors between designs and mainly reused Problem feedback, whereas students may not have noticed or attended to underlying issues between designs. For instance, one student mentioned that they did not use the feedback suggestions because they "*rarely pointed out the same things when critiquing interfaces*."

This exploratory deployment investigated how students reuse feedback and respond to interactive guidance in the classroom. To understand how a system with these features compares to a standard feedback system, the next study was a controlled between-subjects experiment.

EXPERIMENTS: SCAFFOLDING FEEDBACK

Following our deployments, we conducted two empirical studies to investigate the impact of suggestions and guidance on feedback quality.

EXP 1: Do Static Suggestions Improve Feedback?

In an online between-subjects study, 40 undergraduate design students were asked to review three restaurant website homepages using CritiqueKit. The task emulated peer review tasks often required in creative courses. This study's suggestion corpus came from a design feedback task on CrowdCrit [25] and was categorized in the Positive, Problem, and Solution categories. We hypothesized that suggestions and guidance would help reviewers provide more specific and actionable comments.

Method: Reviewing Restaurant Websites

40 participants were randomly assigned to either the CritiqueKit condition or the Control condition (20 in each). CritiqueKit participants used the same version of CritiqueKit as DEP 2 with all features available (Figure 4). Control participants used an otherwise identical version consisting solely of a comment text box. Upon landing on the homepage of either version, participants were provided with a scenario explaining that three restaurant owners are seeking feedback on their new website design. Participants were given a brief tutorial of CritiqueKit's features and an explanation of what makes for good feedback. There were no restrictions or requirements on time spent or amount of feedback to provide. We compared the percentages of comments in five categories. Comments including a supportive element were labeled as *Positive Only* or *Positive+Specific*. Comments including a critical element were labeled *Problem Only*, *Solution Only*, or *Problem+Solution*.

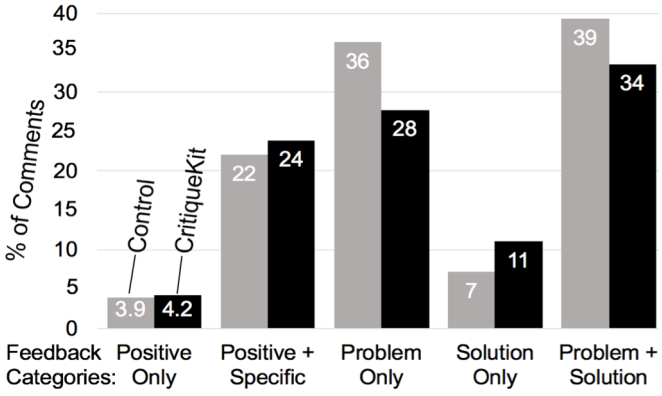

Figure 5. A plurality of feedback in both conditions in EXP 1 identified both a problem and solution (*i.e.*, was actionable). Feedback that was only positive was the rarest. There were no significant differences between conditions for these categories.

Results: Static Suggestions Were Not Helpful

With static suggestions and interactive guidance, there were no significant differences between conditions. (To foreshadow, we will see differences in EXP 2, which adds adaptive suggestions). Participants provided a total of 323 comments (168 for control, 155 for CritiqueKit). The average number of words per comment was not significantly different between conditions (Control: m=29.07, SD=23.64; CritiqueKit: m=23.22, SD=17.3) ($F(1,38)=2.52$, $p=.11$).

Suggestions & Guidance Did Not Affect Type of Feedback

The distribution of the five category types did not vary significantly between conditions ($x^2 = 4.80$, $df = 4$, $p = .31$) (Figure 5). In both groups, participants provided mostly Problem + Solution feedback (39% in Control; 34% in CritiqueKit).

Most CritiqueKit Participants Corrected Category Labels

65% of CritiqueKit participants actively used the guidance panel, making a total of 85 corrections to categories. Interaction with the guidance panel may have indicated attention to the feedback characteristics. As the study was online, we don't know how many of the other 35% were influenced by the guidance panel.

Unfortunately, People also Reused Vague Suggestions

11 distinct suggestions from the corpus were reused. 8 of these were vague; 3 were specific. 15 of 155 reviews included a reused suggestion. This seems especially low given the high engagement with the guidance panel. We see two reasons for this: First, the suggestions came from CrowdCrit [25], where participants provided feedback on a weather app design. The study task was different than the task for which the suggested feedback was originally given, and novices may have had a limited ability to see the deep structure behind a suggestion and reapply it in a new context. Second, the suggestions were created by crowd workers and of uneven quality.

The suggestions selected were typically short, positive comments, perhaps because students did not know how to apply them in the specific context. For example, the most commonly reused suggestion was "great use of color" (reused 3 times). This result is similar to DEP2 in which students did not find feedback provided by other peers or novices to be

useful and generalizable. Feedback suggestions may require more curation or quality control to be most useful.

Suggestions & Guidance Should Work in Concert

While this version of CritiqueKit contained both feedback suggestions and interactive guidance, these features functioned independently. Regardless of the categories checked in the guidance panel, the suggestions remained static and presented in the same order for each participant, potentially making them easy to ignore if they were irrelevant to the context. Participants may have paid attention to only one feature at a time. The next study investigated the question of whether adaptively-presenting feedback suggestions along with interactive guidance improves feedback.

EXP 2: Do Adaptive Suggestions Help?

The second experiment used the final version of CritiqueKit described in the system section to test the hypothesis that adaptively-presented suggestions combined with guidance would improve feedback by increasing the fraction of feedback that is specific, actionable, and/or justified.

Method: Reviewing Paper Prototypes

We conducted a between-subjects in-person study with 47 (27 female) participants. Participants were recruited from an undergraduate subject pool within the Psychology and Cognitive Science departments. Participants were randomly assigned to either the CritiqueKit (n=24) or Control (n=23) conditions. 44 of these participants had no design course experience; 3 participants had taken at least one design course. 28 spoke English as a second language.

Participants were asked to provide feedback on two designs from students enrolled in an online course who volunteered to receive more feedback on their work. These designs were PDFs of mobile application paper prototypes. The review criteria included whether the prototype supported the student's point of view and whether it seemed easily navigable. Participants were first shown the design instructions and review criteria and then given a short tutorial of CritiqueKit as well as an explanation of what makes for good feedback. CritiqueKit participants had all features of CritiqueKit available to them (Figure 2), while Control participants used a version that consisted of only a textbox for their feedback. The task took about 30 minutes to complete. After providing feedback on both designs, participants were interviewed about their feedback process and use of CritiqueKit.

Presenting Feedback Suggestions Adaptively

The categories on the guidance panel and their definition used for coding participants' responses were the following:

Specific: relates directly to the review criteria

Actionable: gives a concrete suggestion for improvement

Justified: provides a reason for why something is done well or should be improved

For DEP 2 and EXP 1, the guidance panel categories sought to encourage specific and actionable feedback (Figure 4). Examining the feedback from our previous studies, we found that "Is Positive" and "Identifies a Problem" did not provide significant guidance as reviewers were generally aware of whether their feedback was positive or critical. In addition, the guidance panel did not explicitly check for justification of feedback. For EXP 2, we revised the categories to "Is Specific," "Is Actionable," and "Is Justified" to also encourage the explanation or reasoning behind feedback. As described in the system section, the checkboxes update as the reviewer types to reflect the categories present in their comment, and the suggestions adapt to show feedback examples from categories not yet present in the comment.

Results: CritiqueKit Participants Provided More Specific, Actionable, & Justified Feedback

Participants provided 158 total comments (79 control, 79 CritiqueKit). The percentage of comments that contained all three categories (specific, actionable, and justified) was significantly higher in the CritiqueKit condition (53%) than in Control (30%) (x^2=8.33, df=1, p=.01) (Figure 1). As an example, this comment meets all three: "The 'more questionnaires' section (Specific) should be made smaller (Actionable) because it is not the main focus of the page." (Justified). The system's heuristic for checking specificity of a comment was quite simple: five words or greater in length. Feedback raters blind to each condition used a more sophisticated and holistic assessment, taking specific to also mean related to the review criteria. With this assessment, 98% of CritiqueKit comments were labeled by raters as specific whereas only 83% of Control comments were. These raters also rated comments from EXP 1 within the specific, actionable, and justified categories to provide a comparison between the two experiments. Interestingly, the percentage of comments containing all attributes in the Control condition was relatively consistent between EXP 1 (35%) and EXP 2 (30%). The percentage of comments with all attributes in the CritiqueKit condition greatly increased between the two experiments (26% versus 53%). Having the checkboxes may have explicitly reminded CritiqueKit participants to ensure their comments satisfy the specific, actionable, and justified categories.

Because longer comments were more likely to contain all three categories, each comment was also scored on a point scale and averaged per participant. Comments received one point for each specific, actionable, and justified idea (Figure 6). A MANOVA with category points as dependent variables shows a significant difference between conditions($F(1,3)=3.21$, $p < .005$). CritiqueKit participants provided more specific ideas than Control participants (Control m= 3.87, CritiqueKit

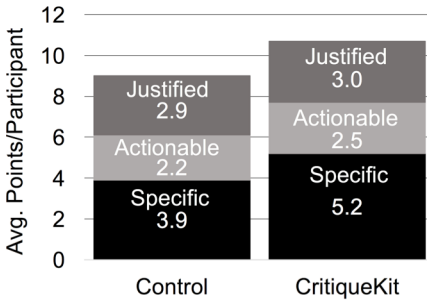

Figure 6. CritiqueKit participants provided more *specific* and *all-three* category ideas that than control participants.

m=5.17, F(1,156)=14.04, $p<.05$). This may be because the suggestions provided examples of relevant ideas and led CritiqueKit participants to address more. CritiqueKit participants also provided more ideas that fit all three categories than control (Control m = 1.0, CritiqueKit m = 2.2, F(1,156)=8.78, $p<.005$). Given that most participants did not have any design experience, the combination of adaptive suggestions and guidance may have been most useful for these reviewers. The suggestions may have provided a starting point while the guidance panel helped them understand how to apply the attributes of good feedback. There were no significant differences in the average number of actionable and justified ideas in comments.

On average, Control comments were 39.3 (SD=30.3) words long and 43.7 (SD=31.4) words for CritiqueKit comments. There was no significant difference in comment length (F(1,156)=1.77, $p=.19$). Unfortunately, we don't know how the feedback improved students' work because they received feedback from both Control and CritiqueKit participants. A longitudinal deployment with the final version of CritiqueKit would likely be more useful in determining the helpfulness of feedback.

Suggestions Helped Reviewers Describe Their Thoughts

Participants rated the suggestions as being generally helpful (m=4.29, SD=0.95, 1-5 Likert Scale). When asked to elaborate on their rating, many participants noted that the suggestions helped them describe their thoughts. One participant remarked, "*I was a bit lost at first because I didn't know how to describe my thoughts. The suggestions helped me figure out how I should describe what I was thinking*." Similarly, another mentioned that "*when I [didn't] know how to put my feedback in words, I could look at the suggestions*." Particularly for participants without any design experience, suggestions helped with appropriate language to use in their feedback. For example, one noted that "*seeing actual wording from a designer's point of view was good so you know how to say what you want to say*." Though a few participants did not directly select suggestions, it is likely that they were inspired or influenced by them as they used similar wording in their own comments.

Still, some participants felt the suggestions were too general and not entirely relevant to the specific design they were reviewing. One participant felt constrained by the suggestions, stating that she ignored them because she wanted to write her own opinions. Suggestions seemed most helpful for participants who used them as a starting point for their own thoughts rather than solely relying on them. Participants who simply selected suggestions tended to list issues without adding their own elaboration. This behavior not only led to incomplete feedback, but also produced depersonalized and scattered comments. For example, one comment that solely relied on suggestions reads "User immediately knows the purpose of the prototype. Good use of grid layout to keep items aligned. Icons should be immediately recognizable to the user." A consideration for future work is to develop feedback suggestions tailored to help reviewers provide more cohesive and contextual comments.

Most participants in this experiment did not have any design experience and may have benefited most from the suggestions. Many participants noted using the suggestions as a way to find ideas whereas students with design experience may already have heuristics and processes in mind when providing feedback. Future work should examine how suggestions and guidance might improve feedback for more experienced learners as well.

Interactive Guidance Helped Remind & Focus Reviewers

Participants were mixed on the helpfulness of the guidance panel (m=3.67, SD=1.2, 1-5 Likert Scale). Those who did find it helpful noted that the categories helped guide their feedback process. For instance, one participant noted that he "*went in order of the checkboxes. First, I provided something specific, then something actionable, then justified it.*" Another noted that the categories helped her know whether her feedback was actually useful or helpful, and one noted that the guidance panel *"[made] sure the feedback is complete and not vague*."

Anecdotally, we observed that when participants said the categories were not useful, it was because they believed them to be inaccurate in their classifications. The accuracy (compared to human raters) for the actionable category was 67% and 75% for the justified category. A participant stated that *"[the checkboxes] didn't always check when I thought they should, so I would just do it myself.*" Another thought the checkboxes were "*quick to judge, it felt like it wasn't reading what I was saying.*" Three participants, who were not native English speakers, found the categories confusing because they weren't sure what they meant. Future iterations of CritiqueKit could include the definition of these categories in the prompts to make the meaning clearer. Interestingly, a couple participants noted that they used the categories as reminders rather than for active guidance. For instance, a participant mentioned that though he felt the interactive guidance was not that accurate, *"[the checkboxes] reminded me to make sure my comment contained specific, actionable, and justified parts, so I'd go and reread through my comment*."

Some participants commented on the adaptive presentation of the suggestions with the guidance panel. For one participant, the suggestions helped him understand what the categories meant. He noted, "*The whole actionable and justified thing, I didn't know what that meant, so the suggestions helped with that*." Observations of participants showed that some clicked on the checkboxes simply to see the suggestions under each one. When asked about how useful they found CritiqueKit in general, participants varied widely in their ratings of usefulness. A more precise measure would allow participants to compare across conditions, which was not possible with this between-subjects design.

GENERAL DISCUSSION

This paper empirically investigated two techniques for scaffolding feedback: reusable feedback suggestions and adaptive guidance. This work can extend to a broader domain, highlighting the benefits of adaptive guidance for learning more generally in creativity support interfaces [38]. Here we discuss and synthesize the findings.

Generating Reusable Feedback Suggestions

This work investigated whether suggestions and guidance can scaffold the feedback process. For this strategy to work, an eye towards reuse and adaptive feedback must be adopted. As Schön's argues, experts may be most capable of recognizing common patterns and giving useful feedback [36]. However, while feedback should be specific, underlying concepts can generalize across contexts. In the studies that used expert-generated feedback suggestions (DEP 1 and EXP 2), participants cited the same reason for why the suggestions were useful: as inspiration. Participants reported that the suggestions helped them find words for their thoughts or helped direct their attention to issues they did not originally notice. This suggests that reusable suggestions should focus attention to common issues rather than specific instances. Our approach demonstrates how expertise on creative work can be scaled by providing feedback on a few to apply to many [22]. This extends work on reusable feedback in coding and writing [3,13,15] while keeping the human in the loop, enabling novices to learn and reuse expert insights.

It is possible that more general suggestions can lead to less personalized feedback, particularly in abstract domains like visual design. We observed this in 7 of the 79 comments from the CritiqueKit condition in EXP 2, in which the four participants simply selected suggestions without further elaboration. A consideration for creating and presenting reusable suggestions is how these suggestions can be both general yet personal to be more helpful to the recipient.

What is the Best Way to Guide Feedback?

Prior empirical work on feedback (*e.g.,* Kulkarni *et al*. [22] and Krause *et al*. [19]) has not compared static and adaptive suggestions. In this paper, we found that people rarely used static suggestions and did not find them helpful; adaptive suggestions were used more and found more helpful. This reinforces prior work demonstrating that adaptive presentation of examples can improve learning [23,37]. By presenting feedback suggestions that directly addressed missing characteristics of a reviewer's feedback, reviewers were prompted on where they could specifically improve, and explicitly shown examples of how to do so.

The second experiment adapted feedback suggestions based on whether their feedback was categorized as specific, actionable, and/or justified. Though some of the prototype's categorizations were misleading or inaccurate (for example, the comment "user flow is simple" was categorized as "Is Actionable" because of the word "use", even though it lacks a concrete suggestion), participants still referenced the three categories when composing their comments. The guidance panel was useful as a reminder to include the attributes of good feedback in their comments. A more sophisticated method for categorization would likely be helpful, though our naïve approach performed reasonably well overall.

The guidance panel focused on three important attributes of good feedback. A consideration is to also provide guidance for emotional content in feedback, as emotional regulation is important to how learners perceive feedback [19,44]. In addition, other characteristics may also contribute to perceived helpfulness, such as complexity or novelty [19], that could be further explored through adaptive guidance.

Creating Adaptive Feedback Interfaces

In order for adaptive guidance to be most effective, the interface should be suitable for adaptation. In the two deployments and first experiment, the suggestions were not curated in any way: more than 1,400 comments were supplied as suggestions, but only 76 of these were reused by reviewers. Having more suggestions available was not beneficial because the suggestions were not sufficiently adaptable and were potentially irrelevant and difficult to browse. EXP 2 introduced a curated approach: experts provided the suggestions with generalizability in mind. Of the 47 suggestions created, 29 were reused. Though fewer suggestions were available, they were more general and adaptable, potentially making them more useful.

Suggestion presentation shares many properties with search interfaces. Like with search, a good result needs to not only be in the set, but toward the top of the set [16]. The second experiment contained fewer suggestions, enabling easier search and browsing. Effective curation and display of suggestions should take into consideration the quality of feedback suggestions and how likely they are to be selected, potentially using frequency or some measure of generalizability as a signal.

CONCLUSION

Looking across the deployments and experiments, adaptive suggestions and interactive guidance significantly improved feedback while static suggestions did not offer significant improvements. These techniques were embodied in the CritiqueKit system, used by 95 feedback providers and 336 recipients. Future work should examine applying other attributes of helpful feedback and further investigate how best to create, curate, and display adaptive suggestions.

Much knowledge work features both underlying principles and context-specific knowledge of when and how to apply these principles. Potentially applicable feedback and review areas include domains as disparate as hiring and employee reviews, code reviews, product reviews, and reviews of academic papers, screenplays, business plans, and any other domain that blends context-specific creative choices with common genre structures. We hope that creativity support tools of all stripes will find value in the ideas and results presented here.

ACKNOWLEDGEMENTS

We thank Kandarp Khandwala and Janet Johnson for help rating feedback. This research was funded in part by Adobe Research.

REFERENCES

1. Heidi Andrade. 2005. Teaching with Rubrics: The Good, the Bad, and the Ugly. *College Teaching* 53, 1: 27–30.
2. Benjamin S. Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13, 6:4–16.
3. Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and Correct: Using Clusters to Grade Short Answers at Scale. In *Proceedings of the first ACM conference on Learning@ scale conference*.
4. Kwangsu Cho, Christian D Schunn, and Davida Charney. 2006. Commenting on Writing Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication* 23, 3: 260–294.
5. Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott 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.
6. Steven P Dow, Anand Kulkarni, Scott R Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM conference on Computer Supported Cooperative Work*, 1013–1022.
7. Ethan Fast, Colleen Lee, Alex Aiken, Michael S Bernstein, Daphne Koller, and Eric Smith. 2013. Crowd-scale Interactive Formal Reasoning and Analytics. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*.
8. Graham Gibbs and Claire Simpson. 2004. Conditions under which assessment supports students' learning Learning and Teaching in Higher Education. *Learning and Teaching in Higher Education* 1, 1: 3–31.
9. Sarah Gielen, Elien Peeters, Filip Dochy, Patrick Onghena, and Katrien Struyven. 2010. Improving the effectiveness of peer feedback for learning. *Learning and Instruction* 20, 4: 304–315.
10. Elena L. Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. 2015. OverCode: Visualizing Variation in Student Solutions to Programming Problems at Scale. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 2.
11. Elena L Glassman, Aaron Lin, Carrie J Cai, and Robert C Miller. 2016. Learnersourcing Personalized Hints. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.
12. Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the ACM SIGCHI Conference on Creativity and Cognition*.
13. Björn Hartmann, Daniel Macdougall, Joel Brandt, and Scott R Klemmer. 2010. What Would Other Programmers Do? Suggesting Solutions to Error Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
14. John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1: 81–112.
15. Andrew Head, Elena Glassman, Gustavo Soares, Ryo Suzuki, Lucas Figueredo, and Loris D ' Antoni. 2017. Writing Reusable Code Feedback at Scale with Mixed-Initiative Program Synthesis. In *Proceedings of the Fourth ACM Conference on Learning@ Scale*, 89–98.
16. Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.
17. Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
18. David Kelley and Tom Kelley. 2013. *Creative confidence : unleashing the creative potential within us all*. Crown Publishing Group.
19. Markus Krause, Tom Garncarz, Jiaojiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique Style Guide : Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 4627–4639.
20. Chinmay Kulkarni, Michael S Bernstein, and Scott Klemmer. 2015. PeerStudio : Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second ACM Conference on Learning@ Scale*, 75–84.
21. Chinmay Kulkarni, Steven P Dow, and Scott R Klemmer. 2012. Early and Repeated Exposure to Examples Improves Creative Work. In *Proceedings of the Cognitive Science Society*.
22. 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 Transactions on Computer-Human Interaction (TOCHI)* 20, 6: 1–31.
23. Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R Klemmer. 2010. Designing with Interactive Example Galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
24. Lan Li, Xiongyi Liu, and Allen L. Steckelberg. 2010. Assessor or assessee: How student learning improves by

giving and receiving peer feedback. *British Journal of Educational Technology* 41, 3: 525–536.

25. 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. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.*
26. Richard L. Marsh, Joshua D. Landau, and Jason L. Hicks. 1996. How examples may (and may not) constrain creativity. *Memory & Cognition* 24, 5: 669–680.
27. Yang Miao, Richard Badger, and Yu Zhen. 2006. A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing* 15, 3: 179–200.
28. Susanne Narciss and Katja Huth. 2006. Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction* 16, 4: 310–322.
29. Huy Nguyen, Wenting Xiong, and Diane Litman. 2016. Instant Feedback for Increasing the Presence of Solutions in Peer Reviews. In *HLT-NAACL Demos*, 6–10.
30. David J. Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* 31, 2: 199–218.
31. David J Nicol and Debra Macfarlane-dick. 2006. Formative assessment and self-regulated learning : A model and seven principles of good feedback practice . Formative assessment and self-regulated learning : A model and seven principles of good feedback practice . *Studies in Higher Education,* 31, 2: 199–218.
32. Arkalgud Ramaprasad. 1983. On the definition of feedback. *Behavioral Science* 28, 1: 4–13.
33. Brian J. Reiser. 2004. Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *Journal of the Learning Sciences* 13, 3: 273–304.
34. Rod D Roscoe, Laura K Allen, Jennifer L Weston, Scott A Crossley, and Danielle S Mcnamara. 2014. The Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition* 34: 39–59.
35. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instrcutional Science* 18: 119–144.
36. Donald A. Schon. 1984. *The Reflective Practitioner: How Professionals Think In Action*. Basic Books.
37. Amir Shareghi Najar, Antonija Mitrovic, and Bruce M Mclaren. 2014. Adaptive Support versus Alternating Worked Examples and Tutored Problems: Which Leads to Better Learning? *LNCS* 8538: 171–182.
38. Ben Shneiderman. 2007. Creativity Support Tools: Accelerating Discovery and Innovation. *Communications of the ACM* 50, 12.
39. Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. 2017. Gradescope: a Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work. In *Proceedings of the Fourth ACM Conference on Learning@ Scale*, 81–88.
40. Nancy Sommers. 1982. Responding to Student Writing. *College Composition and Communication* 3311062, 2.
41. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right : Testing Many Is Better Than One. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 1243–1252.
42. Keith Topping. 1998. Peer assessment between students in colleges and universities. *Review of Educational Research* 68, 3: 249–276.
43. Sheng-Chau Tseng and Chin-Chung Tsai. 2007. On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education* 49: 1161–1174.
44. Sara Värlander. 2008. The role of students' emotions in formal feedback situations. *Teaching in Higher Education* 13, 2: 145–156.
45. Wenting Xiong, Diane Litman, and Christian D. Schunn. 2012. Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research* 4, 2: 155–176.
46. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*: 1005–1017.
47. 2017. Turnitin Feedback Studio. Retrieved from http://turnitin.com