Worker Expertise and Expert Rubrics in Crowdsourced Design Critique

Alvin Yuan
Kurt Luther
Markus Krause
Steven P. Dow
Björn Hartmann

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-223
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-223.html

December 1, 2015

Acknowledgement

We wish to thank the instructors of the course through which we ran the study, the student participants, and the crowd-workers. Financial support provided by the National Science Foundation under IIS grants 1210836, 1208382, and 1217096, and by the German Academic Exchange Service (DAAD).

Worker Expertise and Expert Rubrics in Crowdsourced Design Critique

by Alvin Yuan

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Björn Hartmann
Research Advisor

(Date)

* * * * * * *

Professor Maneesh Agrawala
Second Reader

(Date)

Worker Expertise and Expert Rubrics in Crowdsourced Design Critique

Alvin Yuan
University of California,
Berkeley
alvin.yuan@berkeley.edu

Kurt Luther
Virginia Polytechnic Institute
and State University
kluther@vt.edu

Markus Krause
Leibniz University
markus@hci.uni-hannover.de

Steven P. Dow
Carnegie Mellon University
spdow@cs.cmu.edu

Björn Hartmann
University of California,
Berkeley
bjoern@eecs.berkeley.edu

ABSTRACT
Expert feedback is valuable but hard to obtain for many designers. Online crowds can provide a source of fast and affordable feedback, but workers may lack relevant domain knowledge and experience. Can expert rubrics address this issue and help novices provide expert-level feedback? To evaluate this, we conducted a 2x2 experiment in which student designers received feedback on a visual design artifact from both experts and novices, who produced feedback using either an expert rubric or no rubric. We find that rubrics help novice workers provide feedback that is rated just as valuable as expert feedback. A follow-up analysis on writing style showed that student designers find feedback most helpful when it is emotional, positive, and specific, and that providing a rubric improves the application of these characteristics in feedback. The analysis also finds that expertise only affects style by increasing critique length, but an informal evaluation suggests that experts may instead produce value through providing clearer justifications.

ACM Classification Keywords
H.5.3. Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces—*Computer-supported cooperative work*

Author Keywords
Design; critique; feedback; crowdsourcing; rubrics.

INTRODUCTION
Feedback has always played an important role in the design process by helping the designer gain insights and improve their work. Designers traditionally receive feedback through studio critique sessions, during which the designers present their work to peers and mentors who then provide comments and suggestions. Unfortunately, replicating this conducive environment can be quite difficult outside of small studio classes. And with the growing demand for design education, designers both inside and outside the classroom will have to find other means of collecting feedback. Some notable online communities exist for this purpose, such as Forrst [43], Photosig [40], and Dribbble [28], but these sources often produce feedback of poor quality and low quantity [40].

The lack of an effective, readily available source of feedback has led some groups to explore crowdsourcing as a potential solution [26, 41]. Crowdsourcing feedback can be appealing due to its scalability, accessibility, and affordability, but it also poses a significant issue in that crowd-workers typically do not possess knowledge or practice in the task domain. To combat this, some strategies include breaking down the work into simpler tasks [1, 41] or providing a rubric to assist the crowd-workers [9, 26].

The goal of these crowd-based systems is ultimately to provide high quality feedback. In that sense, it can often be useful to compare the feedback produced by these systems to feedback produced by experts. Experts have years of domain knowledge and practice, which can enable them to identify more issues, provide more comments on those issues, and suggest more specific changes [5].

Recent studies have explored the value of crowd feedback systems [27, 42, 16], but they have yet to directly evaluate the feedback produced by these systems against expert feedback. As a result, it remains an open question the degree to which these systems are effective at producing valuable feedback relative to experts. We supplement previous work by exploring this question. In particular, we evaluate the effectiveness of providing rubrics to novice crowd-workers by comparing the perceived value of feedback they produce to the perceived value of feedback produced by experts.

To this end, we conducted an experiment in which students from a design class submitted designs and received feedback from both novices and experts. Workers produced feedback using one of two workflows: one provides a rubric of de-

Submitted to CSCW 2016.

sign principles while the other does not. Students then rated the helpfulness of each critique they received. We find that without rubrics, experts provided more helpful feedback than novice workers. However, the addition of expert rubrics improved the novices' performance to the point that it was not statistically different from that of experts.

We then ran a linguistic analysis on the writing style of the critiques to try to uncover the features that students find helpful in feedback. We found evidence that critique length, emotional content, language specificity, and sentence mood all correlate with higher ratings. We also found that providing rubrics led to better application of these features in the feedback presented to designers. Together, these results suggest that writing style matters in feedback and that rubrics help improve the quality of feedback by improving writing style.

Expertise, however, only correlated with critique length. This suggests that experts produce valuable feedback through means which are not explained by writing style alone. We briefly investigate this by qualitatively comparing feedback from experts with no rubrics and novices with rubrics. We coded the highest rated critiques of each group as either containing a strong justification, a weak justification, or no justification and find that highly rated feedback from experts often contain clear justifications of the issues and suggestions being presented. On the other hand, the justifications provided by novices often seem shallow and loosely related to their respective issues and suggestions. Thus, the value of experts may lie in their ability to clearly explain the points they raise when providing feedback.

RELATED WORK

The Importance of Feedback

Developing almost any skill generally requires both practice and feedback [30]. Feedback in particular helps the recipient develop a better understanding of the goals or qualities of standard, how the recipient is progressing towards those goals, and what can be done to progress even more [18]. It accomplishes this by helping the recipients refine "information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" [38].

Within the context of creative design, feedback plays a central role, as it helps guide designers towards their next iteration in the design process [10]. It helps the designer understand design principles [13], recognize how others perceive their work [23], and explore and compare alternatives [7, 35]. Thus, there can be a lot of value in making feedback accessible to a wide range of designers.

Sources of Feedback

The most common sources of feedback are instructors and peers. In standard classroom settings, instructors provide feedback by writing comments on drafts or proposals and by grading assignments. Peer feedback generally involves students from the same class inspecting each other's work and has been employed successfully in many contexts including design [6, 34], programming [4], and essays [37]. Feedback through self-assessment has also been explored for writing consumer reviews, achieving comparable results to external sources of feedback [9]. Additionally, automated feedback has been applied in some contexts such as essay grading [19].

Design feedback typically takes place in the form of a studio critique. During these sessions, designers first present their work, then members of the studio, peers and instructors, provide feedback to help improve the design. Studio critique is an effective method for delivering design feedback [32], but it doesn't scale well and is not generally available to many designers.

Alternatively, some online communities such as Forrst [43], Photosig [40], and Dribbble [28] exist where people can mutually provide feedback on each other's designs, but often these produce sparse, superficial comments [40]. Novices in such communities also often experience evaluation apprehension and may be hesitant to share preliminary work [28].

Crowdsourcing Design Feedback

Recently, crowdsourcing has also been explored as another potential avenue for collecting feedback. Crowdsourcing feedback is particularly appealing due to its scalability and accessibility outside of classroom or studio contexts. Crowds are also capable of contributing diverse perspectives that may be difficult to find within a classroom [8]. Some sites such as Five Second Test [36] and Feedback Army [12] use crowds to gather general impressions and reactions to a submitted design, often through having crowd workers provide free-form responses to open-ended questions.

Another set of crowd-based systems aims to provide more structured feedback. Xu et al. presented Voyant, which breaks down the feedback process into smaller tasks involving identifying elements, first noticed elements, and impressions, as well as rating how well goals are communicated and guidelines are followed [41]. Luther et al. presented CrowdCrit, which instead has workers provide critiques and supplies them with scaffolding in the form a rubric of design principles and critique statements [26]. We focus our attention on this latter set of crowd systems, which make use of structure to improve the quality of crowd feedback.

Structuring Crowd Feedback to Match Expert Feedback

Crowd-based systems often have to be conscious of the fact that workers may have little experience in the domain of the task. In the past, such systems have accommodated workers and achieved better results by providing more structure to their tasks. Soylent showed that constraining open-ended tasks and breaking them down into clearly delimited chunks improves the overall quality of work produced by the crowd [1]. Shepherd provided structure in the form of rubrics that helped scaffold and set expectations [9].

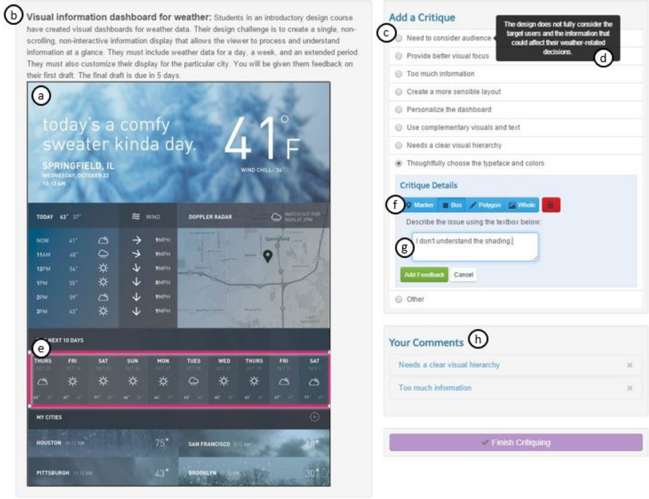
These systems often strive to match the quality of work produced by experts, who are considered to have built a mastery of domain knowledge and performance standards from years of deliberate practice [11]. Experts can also be thought of as having better strategies, knowing which strategies are generally better, being better at choosing strategies, and being better at executing strategies [25, 33]. It seems logical then that experts would be better at providing feedback than non-experts, and empirically experts have been found to produce lengthier comments, produce more idea units, and suggest specific changes more often than their less experienced counterparts when providing feedback on papers [5]. Interestingly, in the context of knowledge transfer and feedback, expertise may have both negative and positive consequences. Experts tend to convey their knowledge more abstractly, which can make it harder for the recipient to immediately understand and apply that knowledge but may also facilitate the transfer of learning to similar tasks [20]. Nevertheless, expert feedback serves as a useful and important baseline to compare results against when determining the effectiveness of feedback.

Voyant and CrowdCrit both use strategies similar to Soylent and Shepherd to structure the design feedback task, and both systems are motivated by the goal of producing higher quality feedback from inexperienced workers. Some recent studies have compared the characteristics of feedback produced by these structured systems against both open-ended feedback and expert feedback with promising results [27, 42, 16], but we have yet to see a study that experimentally evaluates how valuable the feedback produced by these crowd-based systems is compared to feedback produced by experts. This paper strives to supplement existing research and fill in this gap, specifically by quantifying the value of providing expert rubrics and comparing it to the value of expertise.

Assessment and Qualities of Effective Feedback

A variety of methods have been proposed and used to evaluate feedback. Some examples include comparing differences between design iterations [27, 42], comparing crowd feedback to feedback produced by a set of experts [27], measuring post-feedback design quality [7], and collecting designer ratings on the feedback [5]. In our study, we opt for the latter method and have designers rate feedback based on its of perceived helpfulness. Perceived helpfulness is believed to mediate between feedback and later revisions [29], and thus may serve as a strong predictor of future performance. It also has the benefit of having fewer potential confounds when measured in an experiment.

Various explanations have also been proposed to define and understand the qualities that make feedback effective. Sadler argues that effective feedback must help the recipient understand the concept of a standard (conceptual), compare the actual level of performance against this standard (specific), and engage in action that reduces this gap (actionable) [30]. Cho et al. examined the perceived helpfulness of feedback in the context of psychology papers and found that students find feedback more helpful when it suggests a specific change and when it contains positive or encouraging remarks [5]. Xiong and Litman looked at peer feedback for history papers and constructed models using natural language processing to predict perceived helpfulness; they found that lexical features regarding transitions and opinions best predict how helpful students perceive feedback [39]. We employ a similar strategy to explore some of these features in the context of visual design feedback and see how rubrics affect the application of such features.


Visual information dashboard for weather: Students in an introductory design course have created visual dashboards for weather data. Their design challenge is to create a single, non-scrolling, non-interactive information display that allows the viewer to process and understand information at a glance. They must include weather data for a day, a week, and an extended period. They must also customize their display for the particular city. You will be given them feedback on their first draft. The final draft is due in 5 days.

Add a Critique

○ Need to consider audience
○ Provide better visual focus
○ Too much information
○ Create a more sensible layout
○ Personalize the dashboard
○ Use complementary visuals and text
○ Needs a clear visual hierarchy
● Thoughtfully choose the typeface and colors

The design does not fully consider the target users and the information that could affect their weather-related decisions.

Critique Details

Describe the issue using the textbox below:

I don't understand the shading.|

○ Other

Your Comments

Needs a clear visual hierarchy ×

Too much information ×

Figure 1. The structured interface.

RESEARCH QUESTIONS AND HYPOTHESES

This study explores how rubrics affect the way people provide design feedback. It seeks to evaluate how effective novice crowd-workers with rubrics can be in providing feedback compared to experts. Additionally, this study also seeks to uncover relevant features of highly valued feedback and demonstrate that rubrics helps emphasize these features. With that in mind, we explore the following research questions:

1. How does the perceived value of feedback produced by novices with rubrics compare to the perceived value of feedback produced by experts? And do experts also benefit from having rubrics?
2. What are qualities of valuable feedback? And how does providing a rubric affect the occurrence of those qualities?

Our first hypothesis is that novices without rubrics will not produce as valuable feedback as experts due to their lack of proficiency in the domain. We predict the addition of rubrics will compensate for the inexperience and enable novices to provide nearly as helpful feedback as experts. Lastly, we suspect experts will not benefit as much from rubrics because they will already be able to provide helpful feedback on their own.

We also hypothesize that valuable feedback, as suggested by Sadler, incorporates design domain knowledge (conceptual), presents a clear issue in the design (specific), and provides guidance in how to resolve the issue (actionable). We suspect that providing rubrics will significantly increase the frequency of these features, because the rubric attempts to enhance feedback by incorporating conceptual design knowledge into critiques, while encouraging workers to elaborate on the conceptual principle with specific details as well as suggestions.

METHOD

Apparatus

We used the CrowdCrit system [26] to collect feedback in our experiment. The system features two feedback interfaces, one with a rubric and the other with no rubric. The rubric consists of a list of applicable design principles to help workers start off critiques. Workers without a rubric must rely entirely on their own understanding of design to produce critiques.

Interface with Rubric

Figure 1 shows the feedback interface with the rubric present. There are two main sections of the interface: information on the design and the critiquing interface. The design information is comprised of an image of the design (a) as well as some context (b) describing the purpose of the design and experience of the designer. Workers produce critiques through the critiquing interface by first selecting a relevant design principle from the rubric (c). Workers can view descriptions (d) for each principle by mousing over the design principle name. The selected principle forms the basis of the critique they wish to create. They can then provide an annotation (e) using the toolbar (f) to visually indicate what part of the design they are referring to. Additionally, they can provide free-form comments (g) to supplement and elaborate on the critique. Finally, workers can review their work via a list of their produced critiques (h) before submitting.

Interface with No Rubric

This interface is the same as the previous, but provides no principles on which to form the basis of a critique. Instead, workers must rely on the free-form comment box to provide all of the details for their critiques. Workers can still use the annotation toolbar, but are never exposed to the design principles when providing feedback.

Procedure

We recruited 15 students from an undergraduate-level design course at our institution. Each student submitted one design from a course assignment which involved creating a weather UI dashboard. Figure 2 shows all of the submitted designs. Students then received crowd feedback to help them iterate on their designs for a subsequent course assignment.

To generate critiques, we recruited 36 crowd workers of varying design experience, 12 from Odesk and 24 from Mechanical Turk. Workers were then randomly assigned to critique either with or without the aid of a rubric. Odesk workers are typically more skilled and work on longer tasks than Mechanical Turk workers, so we had them critique 8 designs each and compensated them with $30. Mechanical Turk workers critiqued 4 designs each (half of Odesk) and were compensated $3, with the expected rate of pay matching US minimum wage. These numbers ensured that each design received feedback from at least 3 workers in each pool and condition. On average, Odesk workers provided 4.3 critiques per design, and Mechanical Turk workers provided 2.0 critiques per design.

To determine expertise, all workers filled out a questionnaire on their previous design experience, providing information on their design training and work experience. We define experts as workers with both a university degree and work experience in a design field; other workers are referred to as novices. Eleven out of 12 Odesk workers were experts. Only one of 24 mturk workers was an expert, whereas 17 had neither work experience nor education in design. The remaining workers often had some work experience but no degree.

Principle Statement	**Principle Description**
Need to consider audience	The design does not fully consider the target users and the information that could affect their weather-related decisions.
Provide better visual focus	The design lacks a single clear 'point of entry', a visual feature that stands out above all others.
Too much information	Take inventory of the available data and choose to display information that supports the goals of this visual dashboard.
Create a more sensible layout	Information should be placed consistently and organized along a grid to create a sensible layout.
Personalize the dashboard	The design should contain elements that pertain to the particular city, including the name of the city.
Use complementary visuals and text	The design should give viewers an overall visual feel and allow them to learn information from text and graphics.
Needs a clear visual hierarchy	The design should enable a progressive discovery of meaning. There should be layers of importance, where less important information receives less visual prominence.
Thoughtfully choose the typeface and colors	The type and color choices should complement each other and create a consistent theme for the given city.
Other	Freeform critique that does not fit into the other categories.

Table 1. The list of principle statements that comprise the rubric.

The rubric of design principles was provided by the course instructors. See Table 1 for the full list of principles and descriptions. The principles were tailored to the assignment, and closely matched the grading rubric as well as general design principles covered in class.

After all critiques were submitted, the student designers then rated the helpfulness of the CrowdCrit feedback they received on their designs. Critiques were shown one at a time in random order, and students rated their helpfulness on a 1–10 Likert scale (10=best). After rating all their critiques, students could also optionally provide free-form comments on what they found helpful in critiques.

Measures

For our experiment we have two independent variables interpreted as factors with two levels each and one ordinal dependent variable.

Independent Variables

The first factor is worker *expertise* with two levels, *expert* and *novice*. Expert workers have a design degree and have worked as a professional designer.

The second factor is the inclusion of *rubrics* in the feedback interface, again with two levels, *rubric* and *no rubric*. The rubric provides workers with a list of applicable design principles to use as starting points for critiques.

Dependent Variable

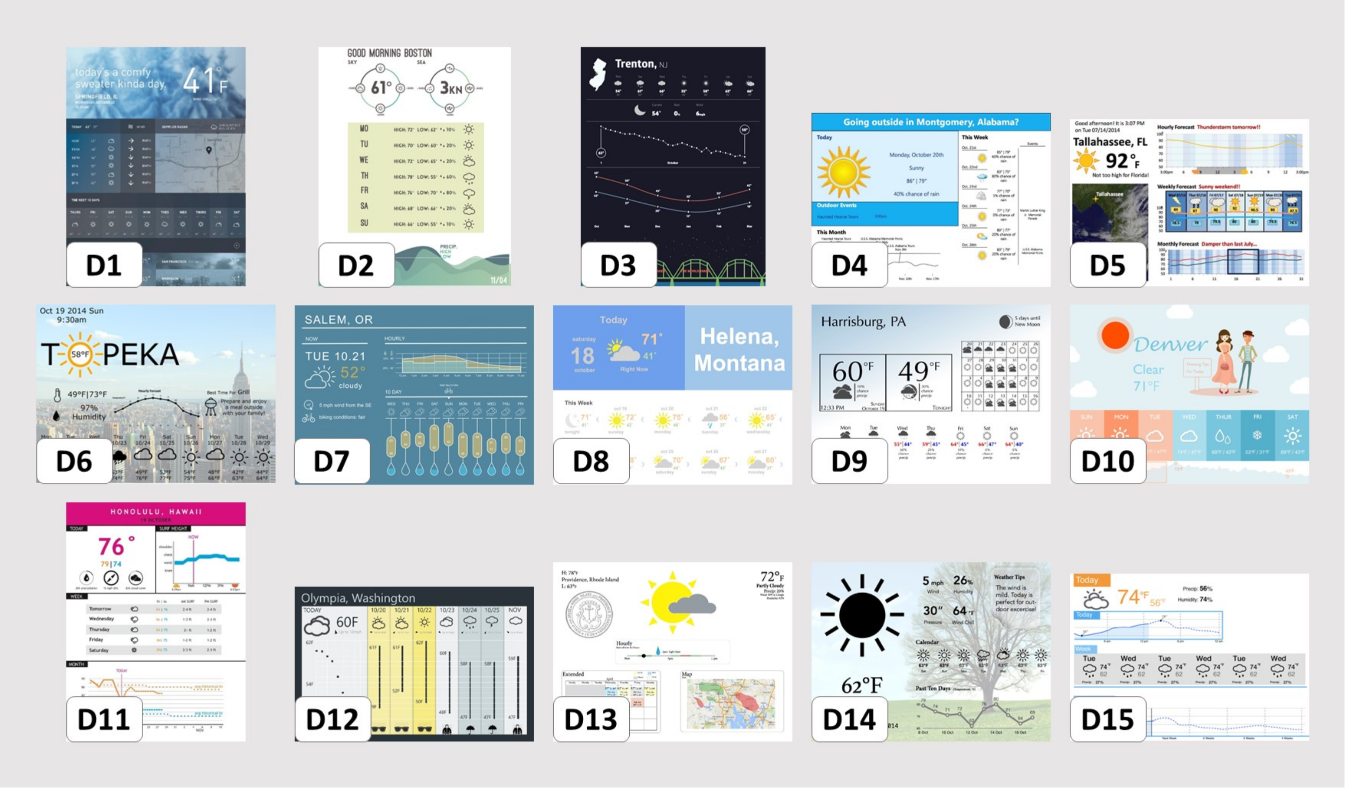


Figure 2. All 15 designs used in the experiment.

The dependent variable is the designer *rating* for each critique, measured using a 1–10 Likert scale. In accordance with [3] we interpret this variable as interval scaled for the purpose of an analysis of variance (ANOVA). Table 2 shows a sample of the low and high rated critiques.

RESULTS

To analyze main and interaction effects of rubrics and worker expertise on ratings, we first conducted a 2-way analysis of variance (ANOVA). In accordance with Harwell [17] and Schmider [31] we assume our sample size n=34 and our substantial effect sizes (Cohens's d>0.6) to be sufficient to meet ANOVA's normality criterion. To ensure equal variance we conducted a Levene's test for homogeneity of variance F(1, 34) = 1.36, p = 0.27. The test does not hold evidence that our data violates the equal variance assumption. The ANOVA showed that rubrics and expertise have significant effects on rating as seen in Table 3. We then conducted a series of post-hoc Welch two sample t-tests using the Holm-Bonferroni correction [21]. From these tests, we find the following results.

Rubrics Help Give Better Feedback

The ANOVA results in Table 3 indicate that rubrics have a positive effect on our rating variable, though the post-hoc Welch two sample t-test is not significant on an Alpha level of 0.05. The difference in ratings between our rubric condition (M = 6.76, SD = 0.70) and our no rubric condition (M = 6.13, SD = 1.7) have the following associated statistics T(25) = 1.95, p = 0.06, d = 0.62.

Experts Provide Better Feedback than Novices

Ignoring rubrics, experts (M = 6.92, SD = 0.61) as expected achieve higher average ratings than novices (M = 6.25, SD = 1.06); t(34) = 2.43, p = 0.05, d = 0.75. The average is almost 10% higher for critiques by experts compared to critiques by novices.

Rubrics Helps Novices More than Experts

We found that novices achieved significantly higher average ratings with rubrics (M = 6.65, SD = 0.647) than without (M = 5.74, SD = 1.28) in our experiment; t(14) = 2.145, p = 0.03, d = 0.89. Rubrics increased the average rating of reviews written by novices by 13.5%. Experts however do not benefit from having rubrics as much as novices: we did not find a significant increase in ratings for experts with rubrics (M = 7.02 SD = 0.79) compared to experts with no rubrics (M = 6.83 SD = 0.41); t(8) = 0.55, p = 0.31, d = 0.32. Figure 3 shows a box plot of the distribution of ratings for all factors and levels.

Highly-Rated Feedback Correlates with Linguistic Features

The first analysis indicates that rubrics have a positive effect on ratings of feedback written by novices. We want to understand what specifically do rubrics provide that lead to these results. To investigate this, we conducted a linguistic analysis with a feature set that has already been used to investigate writing styles in an educational setting [22, 24]. We used the following subset of features: critique length (average word

Low Rated Critiques	High Rated Critiques
Information should be placed consistently and organized along a grid to create a sensible layout. The design is just all over the place. Too many black blocks all over the place. – Novice with rubric to D12, rating=3.	*The type and color choices should complement each other and create a consistent theme for the given city.* The white grid causes some focus issue, it should be darker and blend in better with the backgrounds to create a more natural and polished look. – Novice with rubric to D12, rating=10.
The design should give viewers an overall visual feel and allow them to learn information from text and graphics. This layout is not too please to look at. – Expert with rubric to D4, rating=2.	*Information should be placed consistently and organized along a grid to create a sensible layout.* Because people read left to right it would be more beneficial to place the current temperature (most important) where the eyes first travel. – Expert with rubric to D13, rating=8.
This is not clear – Novice with no rubric to D15, rating=1.	I think this section should be at the top to make it clear that it is the current forecast, as well as looking more visually balanced. – Novice with no rubric to D3, rating=9.
overall this is a great layout – Expert with no rubric to D1, rating=2.	I would suggest putting the actual dates of the weeks here instead of ”3 weeks”. That gives the user less mental work to do to figure out what is in that week. – Expert with no rubric to D15, rating=10.

Table 2. A sample of low and high rated critiques produced by crowd workers. If the rubric was provided, the feedback shown to students includes the selected principle description, shown in italicized text.

Variable	df	SS	MS	F	p	sig.
(R)ubrics	1	3.68	3.68	4.66	0.038	*
(E)xpertise	1	4.22	4.22	5.35	0.027	*
RxE	1	1.01	1.01	1.28	0.265	
Residuals	34	26.0	0.78			

Table 3. ANOVA results of the main and interaction effects of Rubrics and Expertise on perceived feedback quality. Both independent variables are factors with two levels. * indicates significance ($p<0.05$).

length, average sentence length, and number of sentences), emotional content (valence and arousal), language specificity, and sentence mood.

We preprocessed all critiques with the NLTK part-of-speech (POS) tagger [2]. We then filtered stop words and words not in Wordnet [14]. Wordnet is a natural language tool that provides linguistic information on more than 170K words of the English language. We also lemmatized the remaining words to account for different inflections.

We wanted to see if writing style relates to ratings and to rubrics, so we measured the Pearson's product-moment correlation for each of these features with our dependent variable, rating, and with the independent variable, rubrics. The features and results are described next.

Longer Critiques Receive Higher Ratings

The first three features we examined were the average number of letters per word, average number of words per sentence, and number of sentences per critiques. For the average word length we consider only those words that have a Wordnet entry and are not stop words. The sentence length is measured including all words returned by the POS-tagger. The number of sentences per critique includes all sentences found by the POS-tagger. All features positively correlate with higher ratings (r(34) = 0.43, p<0.01, r(34) = 0.49, p<0.01, r(34) = 0.37, p = 0.03). We also found that critiques from the rubric condition have significantly longer words(M = 8.2, SD = 1.7) and sentences(M = 22.4, SD = 3.18) compared to critiques(M = 12.1, SD = 1.7; M = 13.9, SD = 4.8) from the no rubric condition with T(34) = 6.8, p <0.001, d = 2.24 and T(30) = 6.01, p <0.001, d = 2.02. Yet, the number of sentences per critique is not significantly different between the rubric (M = 5.21, SD = 3.89) and no rubric (M = 5.09, SD = 3.89) conditions with t(34) = 0.10, p = 0.46.

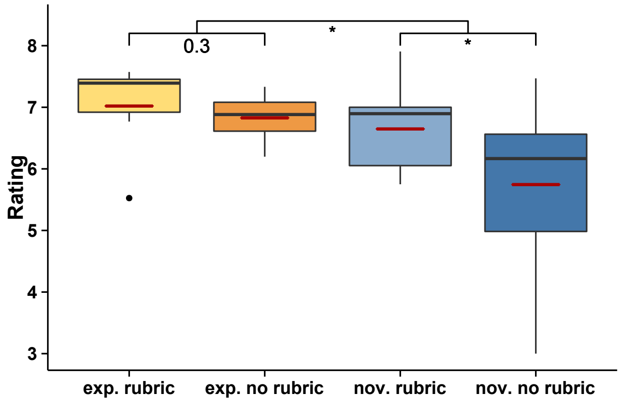


Figure 3. The two left-most columns show the ratings for reviews written by experts (exp.). The two right-most columns show ratings for reviews written by novices (nov.). Results from the rubric condition are shown in lighter colors and marked with (rubric). Results from the no rubric condition are darker and marked with (no rubric). Red lines indicate means. * indicates significance ($p<0.05$).

Emotional Critiques Receive Higher Ratings

The next two features we looked at were valence and arousal. Valence refers to whether the critique is positive, negative, or neutral, and arousal represents how strong the valence is. The normalized value of valence and arousal ranged from -1 to 1 and 0 to 1, respectively. Some examples, with normalized feature values, are provided below.

- Valence=1.0 and arousal=1.0: *This is awesome! I love the map and the hourly weather tool– please keep those!*
- Valence=-0.5 and arousal=0.5: *This graphic is confusing. Is it for show or information? Difficult to tell. Thusly, making the slide hard to read.*
- Valence=0.0 and arousal=0.0: *The fact that it is the same size as the ”sun” has the two elements compete for focus.*

Positively written and emotional critiques received higher average ratings as both, valence and arousal correlate with ratings (r(34) = 0.66, p<0.001 and r(34) = 0.42, p = 0.01). We also found that critiques in the rubric condition have a higher average arousal (M = 0.16, SD = 0.07) and valence (M = 0.82, SD = 0.07) than critiques from the no rubric condition (M = 0.04, SD = 0.15; M = 0.73, SD = 0.09) with T(21) = 2.99, p = 0.003, d = 1.04 and T(31) = 3.07, p = 0.002, d = 1.03 respectively.

Specific Critiques Receive Higher Ratings

Another feature we explored was specificity, which refers to how specific the words in the critique are. We measured specificity by determining how deep each word appears in the Wordnet structure. Words that are closer to the root are more general (e.g. dog) and words deeper in the Wordnet structure are more specific (e.g. labrador). Word depth ranges from 1 to 20 (most specific). To simplify the analysis and presentation we normalize specificity to range from 0.0 to 1.0.

- Specificity=1.0: *This would be good information to include if it had a more unique role such as "Haunted Hearse Tours Today @ 3PM, best to wear a light sweater because it will be sunny but with a light breeze" But because it doesn't serve much of a role directly to the weather display, it is more information to digest and therefore distracting from what you're trying to present to the viewer.*
- Specificity=0.0: *Try using text to indicate what type of information we are looking at.*

Higher specificity correlated with higher ratings (r(34) = 0.63, p<0.001). The average specificity was significantly higher in the rubric condition (M = 0.62, SD = 0.06) than the no rubric condition (M = 0.47, SD = 0.11) with T(25) = 5.06, p<0.001, d = 1.74.

Critiques that Question or Suggest Receive Higher Ratings

The last feature we considered involved looking at the moods of sentences in each critique. Each sentence was classified as either indicative (written as if stating a fact), imperative (expressing a command or suggestion), or subjunctive (exploring hypothetical situations). The feature, which we refer to as active, corresponds to the ratio of non-indicative sentences in a critique, with values falling between 0 and 1. See below for some examples.

- Active=1.0: *I would suggest displaying this information in a more creative manner, or at least using an actual table.*
- Active=0.0: *The text here does not contrast well with the background.*

Active sentences correlated with higher ratings (r(34) = 0.36, p = 0.03). Critiques are significantly more active in the rubric condition (M = 0.66, SD = 0.20) than the no rubric condition (M = 0.38, SD = 0.27) with T(30) = 3.56, p<0.001, d = 1.20.

Expertise Only Correlates with Critique Length

We also examined the correlation between the features and expertise of the worker. We found only one significant correlation between our features and expertise. Only the length of critiques measured in number of sentences is correlated with expertise r(34) = 0.625, p<0.001. We found that experts (M = 8.57, SD = 4.19) have a significantly higher average critique length than novices (M = 3.51, SD = 2.33). Although this difference in means is significant T(14) = 3.8988, p<0.001 we wanted to uncover a little more of what sets expert feedback apart in terms of feedback content. To this end, we examined and compared the highest rated feedback from experts with no rubrics and from novices with rubrics. We coded all critiques rated 9 or 10 from these groups as either having a strong justification, a weak justification, or no justification. We found that the expert feedback more often featured clearer justifications of the issues pointed out and the suggestions proposed. For example, compare the high rated feedback from an expert with no rubric and a novice with rubric in Table 2. The expert feedback explains how using actual dates instead of relative times reduces the mental effort required by the reader. As a result, the designer is able to act on the suggestion with an understanding of why it helps. The novice feedback does provide a justification, but the connection is not immediately obvious. The designer may understand the suggestion proposed and may even be able to act on it, but it is up to the designer's knowledge and experience to understand why such a change would lead to a "more natural and polished look". Among the expert feedback we examined, we find that roughly half of them feature a strong justification. Among the novice feedback, we find only about 20% of them feature a strong justification, though about 67% feature a weak justification. Sometimes the selected principle from the rubric acted as justification, though in these cases it was more often a weak justification. These justifications seem to be the reason why expert feedback is longer, and may also help explain why expert feedback is rated highly.

DISCUSSION

We now revisit our original research questions and discuss our findings from the results.

Research Question 1: Rubrics and Expertise Both Produce Valuable Feedback

First, we found that design experts performed better than novice crowd workers. This is not surprising to see, as experts ought to be better at finding and articulating issues, though it does serve as some validation that the ratings were reasonable.

We also found that rubrics do not provide significant aid to experts. One potential explanation for this is that experts can already recall and apply design principles. They simply might not benefit from having the system present these principles to them. This finding suggests that rubrics may not be necessary in certain contexts. If the feedback providers are expected to be reasonably educated and experienced in the domain, then free-form feedback may be just as effective.

Most importantly, we found that novices with rubrics perform nearly as well as experts (in terms of the perceived value of their critiques), but without rubrics they do significantly worse. This is a good indication that crowd feedback systems can be as effective as experts in producing helpful feedback, and that expert rubrics are an effective method for structuring feedback tasks.

All of these findings together support our original hypothesis regarding the effect of rubrics and expertise. To summarize, experts do not seem to benefit much from having rubrics, but novices perform much better when it is provided. The benefit is significant enough that when given rubrics, novice crowd workers can produce feedback nearly as helpful as feedback from experts. Considering the cost of using a crowd-based system versus the cost of finding and hiring experts, such systems provide a significant and viable opportunity to designers seeking feedback.

However, it is important to keep in mind that these results deal with perceived utility and not actual utility. This study does not show how this feedback translates to actual revisions in the design. It is quite possible that what designers value and what designers use in feedback are two separate notions, and an important next step would be to investigate this.

Research Question 2: Writing Style Matters in Feedback and Rubrics Improve Style.

The latter half of the analysis looked at language features regarding writing style in the feedback text, and found multiple features that positively correlate with ratings. When we considered all possible combinations of the features, we found that the combination of arousal, valence, and specificity in particular achieves the highest correlation with rating. Though only correlational evidence, we interpret this finding to suggest that the application of these features leads to higher ratings. We discuss how this interpretation applies to the individual features next.

Writing Style can Help Direct, Motivate, and Clarify

Arousal indicates a valence, either praise or criticism, and the presence of arousal may make it easier for the designer to interpret a piece of feedback. Negative feedback indicates something to fix and positive feedback indicates something to keep, but neutral feedback may leave the designer without direction. This reasoning overlaps with our interpretation of Sadler's theory that good feedback is actionable. We suspect that the active feature captures a similar quality, which may explain why it did not also contribute to the best combination of features.

The finding that positive valence correlated with higher ratings may be an indication of the conventional wisdom that it is better to point out both positives and negatives rather than being overly critical. As mentioned previously, positive feedback has the virtue of informing the designer what elements are working well and should be kept or even emphasized further. Positive remarks can also be encouraging to the recipient [15], and thus may be considered helpful even in a purely motivational sense.

Specificity is a fairly straightforward feature that also appears in our hypothesis based on Sadler's proposed qualities of good feedback. Specificity aids interpretation by providing concrete details and adding clarity to the focus of the feedback. It also suggests that the feedback provider tailored his comments to the particular design and designer. It seems reasonable that these qualities would improve the perceived helpfulness of the feedback. We suspect that critique length acts as a weaker proxy for specificity, as the inclusion of specific details often involves longer critiques.

Rubrics Improve Feedback By Improving Writing Style

We also found that rubrics help workers improve along all these features. This provides some nice clarity into how and why rubrics are beneficial. In particular, the style in which feedback is written matters to student designers and rubrics help encourage workers to write in a more helpful style. The analysis we conducted does not address feedback content, but investigating this in the future could provide additional insight. It does, however, open up an interesting avenue for research that examines strategies which focus on improving feedback through improving style rather than content.

Justifications Also Matter

An unusual result was that expertise did not correlate with any of the linguistic features in our analysis other than number of sentences. Experts do produce valuable feedback for designers, but the value of their feedback is not adequately explained by writing style. Instead, the value provided by experts may lie in their ability to produce clear justifications of the issues and suggestions they present. These strong justifications lead to more cohesive pieces of feedback which facilitate understanding and applicability. As one designer (D11) adequately put it, ”It was also hard to distinguish taste from objective comments: some people loved the colors, some people hated them. I would've preferred more justification.”

It is not entirely surprising to see this distinction between experts and novices. After all, it is not expected that novices, some of whom have zero design experience, be able to provide clear justifications of their critiques. Additionally, this notion aligns with our hypothesis that good feedback incorporates conceptual knowledge, as justifications are often based on such knowledge. And in fact, the rubric is designed to help compensate for the worker's lack of conceptual knowledge by providing principles to use as justification. The trade-off here though is that the more generally applicable a principle is, the less cohesive it is to any individual piece of feedback. Further investigation can help provide additional insights into the value produced by experts and how to best design systems to replicate that value.

FUTURE WORK

Revisit Effects on Design Iteration

As mentioned before, this study only investigated the effect of rubrics on perceived utility. It still remains to be shown how feedback produced using rubrics compares to both expert feedback and simple open-ended feedback in terms of enabling better redesigns. Some studies have attempted to address this point with mixed results, but no experiment that we know of has demonstrated this claim. However, it is a crucial claim to show and definitely worth investigating.

Further Explore Linguistic Analysis Findings

Our initial work on the linguistic analysis of feedback opens up a few avenues to explore. As mentioned earlier, the analysis only provided correlational evidence, so there is still the question of whether these features have a causal relationship with perceived utility. Another interesting avenue involves exploring systems that structure the feedback task to explicitly improve style. Perhaps the system could predict the perceived value of a potential critique based on these stylistic features and then automatically suggest ways to improve the critique back to the worker. For example, if the piece of feedback is written with a neutral valence (no arousal), the system could suggest to the worker to make it clearer whether he/she is criticizing or praising the design. Such a system may even provide additional benefit by educating crowd workers on how to provide valuable feedback.

Further Analyze Expert Feedback

The linguistic analysis suggests how rubrics might add value to feedback but did not explain how experts produce valuable design feedback. Some initial qualitative analysis suggests one possible explanation, that experts add clear and meaningful justifications to their critiques, leading to more cohesive pieces of feedback. Conducting more investigation on the role of expertise can help provide a deeper understanding of the value of feedback, and this in turn can help motivate new ways of structuring feedback tasks that seek to emulate expert-level feedback.

Investigate Other Forms of Structuring Feedback Tasks

In our experiment, we only tested one system's structure which involved providing rubrics. Other systems use structure in different ways, some of which don't even involve inputting text. Having demonstrated that providing rubrics is indeed beneficial and comparable to hiring experts, a natural followup would be to investigate different strategies of structuring feedback tasks and their trade-offs. This can help deepen our understanding of the role of structure in crowd feedback systems and can lead to leaner and more effective implementations.

CONCLUSION

Crowd feedback systems have the potential to provide high quality feedback to a wide range of designers, but existing research had yet to evaluate their value against the value obtained by hiring experts. We fill in this gap and find evidence that indeed, novice crowd-workers supplied with rubrics can produce just as helpful design feedback as experts. We supplement this finding with additional details as to how rubrics and expertise might be generating value in feedback: rubrics seem to enhance the written style of feedback which student designers find helpful, whereas expertise allows workers to provide stronger, clearer justifications. We hope that our findings motivate further investigation as to how these systems can be designed and utilized best in order to promote widespread accessibility to highly effective feedback.

ACKNOWLEDGMENTS

We wish to thank the instructors of the course through which we ran the study, the student participants, and the crowd-workers. Financial support provided by the National Science Foundation under IIS grants 1210836, 1208382, and 1217096, and by the German Academic Exchange Service (DAAD).

REFERENCES

1. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 313–322. DOI: http://dx.doi.org/10.1145/1866029.1866078

2. Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. DOI: http://dx.doi.org/10.1097/00004770-200204000-00018

3. James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42, 12 (2008), 1150–1152. DOI: http://dx.doi.org/10.1111/j.1365-2923.2008.03172.x

4. Donald Chinn. 2005. Peer Assessment in the Algorithms Course. In *Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE '05)*. ACM, New York, NY, USA, 69–73. DOI: http://dx.doi.org/10.1145/1067445.1067468

5. 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 (July 2006), 260–294. DOI: http://dx.doi.org/10.1177/0741088306289261

6. Barbara De La Harpe, J. Fiona Peterson, Noel Frankham, Robert Zehner, Douglas Neale, Elizabeth Musgrave, and Ruth McDermott. 2009. Assessment Focus in Studio: What is Most Prominent in Architecture, Art and Design? *International Journal of Art & Design Education* 28, 1 (Feb. 2009), 37–51. DOI: http://dx.doi.org/10.1111/j.1476-8070.2009.01591.x

7. Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2807–2816. DOI: http://dx.doi.org/10.1145/1978942.1979359

8. Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 227–236. DOI: http://dx.doi.org/10.1145/2470654.2470686

9. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. DOI: http://dx.doi.org/10.1145/2145204.2145355

10. Steven P. Dow, Kate Heddleston, and Scott R. Klemmer. 2009. The Efficacy of Prototyping Under Time Constraints. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition (C&C '09)*. ACM, New York, NY, USA, 165–174. DOI: http://dx.doi.org/10.1145/1640233.1640260

11. K. Anders Ericsson, Ralf Th Krampe, and Clemens Tesch-romer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* (1993), 363–406.

12. Feedback Army. Website Usability Testing Service - Feedback Army. (2015). http://www.feedbackarmy.com/

13. Edmund Burke Feldman. 1994. *Practical Art Criticism*. Pearson, Englewood Cliffs, N.J.

14. Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

15. Thomas C. Gee. 1972. Students' Responses to Teacher Comments. *Research in the Teaching of English* 6, 2 (Oct. 1972), 212–221. http://www.jstor.org/stable/40170807

16. 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 ACM Creativity & Cognition 2015*. ACM, Glasgow, Scotland.

17. M. R. Harwell, E. N. Rubinstein, W. S. Hayes, and C. C. Olds. Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. (1992). DOI: http://dx.doi.org/10.3102/10769986017004315

18. John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (March 2007), 81–112. DOI: http://dx.doi.org/10.3102/003465430298487

19. M.A. Hearst. 2000. The debate on automated essay grading. *IEEE Intelligent Systems and their Applications* 15, 5 (Sept. 2000), 22–37. DOI: http://dx.doi.org/10.1109/5254.889104

20. Pamela J. Hinds, Michael Patterson, and Jeffrey Pfeffer. 2001. Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *Journal of Applied Psychology* 86, 6 (2001), 1232–1243. DOI: http://dx.doi.org/10.1037/0021-9010.86.6.1232

21. Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

22. Niklas Kilian, Markus Krause, Nina Runge, and Jan Smeddinck. 2012. Predicting Crowd-based Translation Quality with Language-independent Feature Vectors. In *HComp'12 Proceedings of the AAAI Workshop on Human Computation*. AAAI Press, Toronto, ON, Canada, 114–115. http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/viewPDFInterstitial/5237/5611

23. Scott R. Klemmer, Björn Hartmann, and Leila Takayama. 2006. How Bodies Matter: Five Themes for Interaction Design. In *Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06)*. ACM, New York, NY, USA, 140–149. DOI: http://dx.doi.org/10.1145/1142405.1142429

24. Markus Krause. 2014. A behavioral biometrics based authentication method for MOOC's that is robust against imitation attempts. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. ACM Press, Atlanta, GA, USA, 201–202. DOI: http://dx.doi.org/10.1145/2556325.2567881

25. P. Lemaire and R. S. Siegler. 1995. Four aspects of strategic change: contributions to children's learning of multiplication. *Journal of Experimental Psychology. General* 124, 1 (March 1995), 83–97.

26. Kurt Luther, Amy Pavel, Wei Wu, Jari-lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2014. CrowdCrit: Crowdsourcing and Aggregating Visual Design Critique. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW Companion '14)*. ACM, New York, NY, USA, 21–24. DOI: http://dx.doi.org/10.1145/2556420.2556788

27. 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 (CSCW '15)*. ACM, New York, NY, USA, 473–485. DOI:http://dx.doi.org/10.1145/2675133.2675283

28. Jennifer Marlow and Laura Dabbish. 2014. From Rookie to All-star: Professional Development in a Graphic Design Social Networking Site. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 922–933. DOI: http://dx.doi.org/10.1145/2531602.2531651

29. Mary L. Rucker and Stephanie Thomson. 2003. Assessing Student Learning Outcomes: An Investigation of the Relationship among Feedback Measures. *College Student Journal* 37, 3 (Sept. 2003), 400.

30. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (June 1989), 119–144. DOI: http://dx.doi.org/10.1007/BF00117714

31. Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. Is It Really Robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology* 6, 4 (2010), 147–151. DOI: http://dx.doi.org/10.1027/1614-2241/a000016

32. Donald A. Schön. 1985. *The Design Studio: An Exploration of Its Traditions and Potentials*. Riba-Publ.

33. Christian D. Schunn, Mark U. McGregor, and Lelyn D. Saner. 2005. Expertise in ill-defined problem-solving domains as effective strategy use. *Memory & Cognition* 33, 8 (Dec. 2005), 1377–1387.

34. 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 (2013), 29. http://www.ieeetclt.org/issues/january2013/Tinapple.pdf

35. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1243–1252. DOI: http://dx.doi.org/10.1145/1124772.1124960

36. UsabilityHub. Five Second Test. (2015). http://fivesecondtest.com/

37. Anne Venables and Raymond Summit. 2003. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* 40, 3 (Aug. 2003), 281–290. DOI: http://dx.doi.org/10.1080/1470329032000103816

38. P.H. Winne and D. L. Butler. 1994. Student cognition in learning from teaching. In *International encyclopaedia of education* (2 ed.), T. Husen and T. Postlewaite (Eds.). Pergamon, Oxford, UK, 5738–5745.

39. Wenting Xiong and Diane J. Litman. 2011. Understanding Differences in Perceived Peer-Review Helpfulness using Natural Language Processing. In *IUNLPBEA '11 Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Stroudsburg, PA, USA, 10–19. http://dl.acm.org/citation.cfm?id=2043132&picked=prox

40. Anbang Xu and Brian Bailey. 2012. What Do You Think?: A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 295–304. DOI: http://dx.doi.org/10.1145/2145204.2145252

41. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1433–1444. DOI: http://dx.doi.org/10.1145/2531602.2531604

42. Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. 2015. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1637–1648. DOI:http://dx.doi.org/10.1145/2675133.2675140

43. ZURB. Forrst. (2015). http://zurb.com/forrst