Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques

Alvin Yuan
UC Berkeley
alvin.yuan@berkeley.edu

Kurt Luther
Virginia Tech
kluther@vt.edu

Markus Krause
ICSI, UC Berkeley
markus@icsi.berkeley.edu

Sophie Vennix
Carnegie Mellon University
siv@andrew.cmu.edu

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

Björn Hartmann
UC Berkeley
bjoern@eecs.berkeley.edu

ABSTRACT
Expert feedback is valuable but hard to obtain for many designers. Online crowds can provide 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 an experiment with a 2×2 factorial design. Student designers received feedback on a visual design from both experts and novices, who produced feedback using either an expert rubric or no rubric. We found that rubrics helped novice workers provide feedback that was rated nearly as valuable as expert feedback. A follow-up analysis on writing style showed that student designers found feedback most helpful when it was emotionally positive and specific, and that a rubric increased the occurrence of these characteristics in feedback. The analysis also found that expertise correlated with longer critiques, but not the other favorable characteristics. An informal evaluation indicates that experts may instead have produced value by 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; expertise; 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, where they present their work to peers and mentors who provide comments and suggestions. Unfortunately, replicating this conducive environment outside of small studio classes can be quite difficult. With the demand for design education growing, designers both inside and outside the classroom must find other means of collecting feedback. Some notable online communities exist for this purpose, such as Forrst [52], Photosig [48], and Dribbble [31], but these sources often produce feedback of poor quality and low quantity [48].

The lack of an effective, readily available source of feedback has led some researchers to explore crowdsourcing as a potential solution [30, 49]. Crowdsourcing feedback can be appealing due to its scalability, availability, and affordability, but it also poses a significant challenge: crowd workers typically do not possess knowledge or skills in specialized task domains. To combat this, some crowd-based systems break down work into simpler tasks (e.g. [1]) or provide rubrics to workers (e.g. [9]). In the domain of design critique, researchers have applied similar strategies to help novice crowds provide feedback more like experts [49, 30, 18]. While prior work demonstrates the plausibility of obtaining relevant and rapid crowd feedback, this paper focuses on the salient differences between expert and novice feedback providers. Almost by definition, experts know more about a domain, but do they provide better feedback? And if so, what characteristics make expert feedback better than novice feedback? Understanding these characteristics can inform the design of technologies to scaffold novice feedback providers and to increase the availability of valuable design feedback.

We investigate the value, specifically the perceived helpfulness, of novice feedback relative to expert feedback, either with or without an expert rubric. We conducted a 2×2 between-subjects experiment where students from a university-level visual design course submitted drafts and received feedback. Novice and expert workers hired from Amazon Mechanical Turk and Upwork produced feedback using one of two workflows: one provides structure using a rubric of design principles, and the other simply asks for open-ended responses. The students, blind to condition, rated the helpfulness of each critique they received. We found that without rubrics, expert feedback was perceived to be more helpful than feedback from novices. However, the addition of rubrics

CSCW '16, February 27 – March 02, 2016, San Francisco, CA, USA
ACM 978-1-4503-3592-8/16/02.
http://dx.doi.org/10.1145/2818048.2819953

improved the perceived value of novice feedback to the point that it was not statistically different from that of experts.

To identify the features that students found most helpful, we conducted a linguistic analysis on the writing style of the critiques. We found evidence that critique length, emotional content, language specificity, and the grammatical mood of sentences all correlate with higher ratings. We also found that providing rubrics led to more occurrences of these features in the feedback presented to student designers. Together, these results suggest that writing style affects the perceived value of feedback and that rubrics can help improve the writing style.

Our model shows that expertise only correlates with critique length and not with other favorable characteristics from our linguistic model. This suggests that the perceived value of expert feedback cannot explained by writing style alone. We investigate the value of expert critique by qualitatively comparing highly-rated feedback from experts without rubrics and novices with rubrics. We coded critique statements from each group and found that highly-rated expert feedback more often contained clear justifications for the issues and suggestions they raise. On the other hand, the justifications provided by novices tended to be shallow and less related to their respective issues and suggestions. Thus, the value of expertise may lie in the ability to clearly explain the rationale behind the feedback. Future work may further explore the qualities of expert feedback and motivate more ways of structuring design feedback tasks to produce high-quality feedback.

RELATED WORK

The Importance of Feedback

Developing almost any skill generally requires both practice and feedback [36]. Feedback helps the recipient develop a better understanding of the goals or qualities of a standard and track how they are progressing towards those goals [20]. 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" [46].

In 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 [14], recognize how others perceive their work [25], and explore and compare alternatives [7, 42]. As digital tools bring design capabilities to an increasingly broad segment of society, there is great potential value in making high-quality feedback available 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. It has been employed successfully in many contexts including design [6, 41, 27], programming [4], and essays [45]. Self-assessment has also proven useful, achieving comparable results to external sources of feedback [9]. Additionally, automated feedback has been applied in some contexts such as essay grading [21] and kitchen design [16].

Design feedback typically takes the form of a studio critique. During these sessions, designers first present their work and the rationale for their work, then peers and instructors provide feedback to help the designer consider how to improve the design. Studio critique is an effective method for delivering design feedback [38], but it does not scale effectively and remains unavailable to many designers.

Some online communities such as Forrst [52], Photosig [48], and Dribbble [31] enable people to mutually provide feedback on each other's designs, but often these produce sparse, superficial comments [48]. Novices in such communities also often experience evaluation apprehension and may be hesitant to share preliminary work [31].

Crowdsourcing Design Feedback

Recently, researchers have explored crowdsourcing as another potential avenue for collecting feedback. Crowd feedback is particularly appealing due to its scalability and availability outside of classroom or studio contexts. This feedback does not necessarily be textual but can also be visual [34]. Crowds are also capable of contributing diverse perspectives that may be difficult to find within a classroom [8]. However, online crowds often fail to pay attention to task details and may also lack domain expertise. Prior work has contributed screening processes to disqualify workers that lack conscientiousness [11] and incentive mechanisms such as the Bayesian Truth Serum [40] to increase work quality. Current commercial design feedback systems sidestep such issues by only eliciting very general impressions and reactions to a submitted design (e.g., Five Second Test [44] and Feedback Army [13]).

Another set of crowd-based systems aims to provide more structured design feedback. Voyant [49] breaks down the feedback process into micro-tasks that involve identifying first-noticed elements, sharing impressions, and judging how well the designer reaches their goals and follows visual design guidelines. CrowdCrit [30] takes a different approach in which workers use a rubric of design principles and critique statements. We focus our attention on this latter set of crowd systems, which make use of rubrics to improve the quality of crowd feedback.

Structuring Crowd Feedback to Match Expert Feedback

Crowd-based systems often have to account for the fact that workers may have little experience in the task domain. 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]. Systems such as Shepherd [9] and PeerStudio [27] provided structure in the form of rubrics that helped scaffold and set expectations. Motif [24] illustrates that scaffolds extracted from expert examples help novices creating video stories.

These systems often strive to match the quality of work produced by experts who have mastered domain knowledge and helped establish standards through years of deliberate practice [12]. Experts tend to develop better strategies and sharper intuition for when to select and how to execute these strategies [29, 39]. It might follow that experts would be better at providing feedback than novices; in fact, experts have been shown to produce longer comments, generate more idea units, and suggest specific changes more often than their less experienced counterparts when providing feedback on writing [5]. However, experts tend to convey their knowledge more abstractly, which may facilitate learning transfer to similar tasks, but can make it harder for the recipient to immediately understand and apply that knowledge [22]. Nevertheless, expert feedback serves as a useful and important baseline to compare results against when determining the effectiveness of feedback rubrics.

Voyant and CrowdCrit use similar strategies to structure design feedback tasks for online crowds. Both systems are motivated by the goal of producing higher quality feedback from inexperienced workers. Recent studies compare the *characteristics* of feedback produced by these structured systems against both open-ended feedback and expert feedback [30, 50, 18]. However, prior research has not empirically investigated the *perceived value* of feedback produced by novices on crowd-based systems compared to feedback produced by experts. This paper builds on prior work by comparing how students value feedback from experts versus novices with expert rubrics, and by conducting language analyses on the feedback to understand what students find valuable.

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 [30, 50], contrasting with feedback produced by experts [30, 27], measuring post-feedback design quality [7], and collecting designer ratings on the helpfulness of feedback [5].

While measuring the impact on design outcomes provides a compelling and naturalistic method to evaluate the effects of feedback, it can be difficult to measure [30] and faces a number of cofounding factors. Designers may lack the ability and motivation to execute changes suggested by feedback. In many design contexts, including the classroom setting for our study, designers receive feedback from many sources such as peers and instructors, making it difficult to attribute design changes to specific feedback sources.

In our study, we opt to evaluate the perceived helpfulness of feedback. Perceived helpfulness directly captures the value of feedback for its recipient and potentially mediates the interaction between feedback and later revisions [35], and thus may serve as a strong predictor of future performance.

Various explanations have been proposed to define and understand the qualities that make feedback effective. Sadler [36] 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). Cho et al. [5] examined the perceived helpfulness of feedback in the context of writing psychology papers and found that students find feedback more helpful when it suggests a specific change and when it contains positive or encouraging remarks.

Xiong and Litman [47] 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. We employ a similar strategy to explore some of these features in the context of visual design feedback and see how rubrics affect the occurrence of such features.

RESEARCH QUESTIONS AND HYPOTHESES

This study explores how rubrics affect the way people provide design feedback. It seeks to evaluate the perceived value of feedback from novice crowd workers with rubrics compared to experts. Additionally, this study also seeks to uncover relevant features of highly valued feedback and investigate how and if rubrics help emphasize these features. With these ideas 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 feedback as valuable as experts due to their lack of proficiency in the domain. We predict that the addition of rubrics will partly compensate for the inexperience and enable novices to provide feedback nearly as helpful as experts. 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 incorporates the qualities suggested by Sadler [36] and Cho et al. [5]. That is, valuable feedback is *conceptual* in that it incorporates design domain knowledge, *specific* in that it presents a clear issue, *actionable* in that it provides guidance on how to resolve the issue, and *positive* in that it also encourages the recipient.

We suspect that providing rubrics will significantly increase the frequency of these features. Rubrics may enhance feedback by incorporating conceptual design knowledge into critiques and encouraging workers to elaborate with specific details and suggestions. They may also draw attention to elements of the design that align well with the rubric principles and give workers the language to describe those successes.

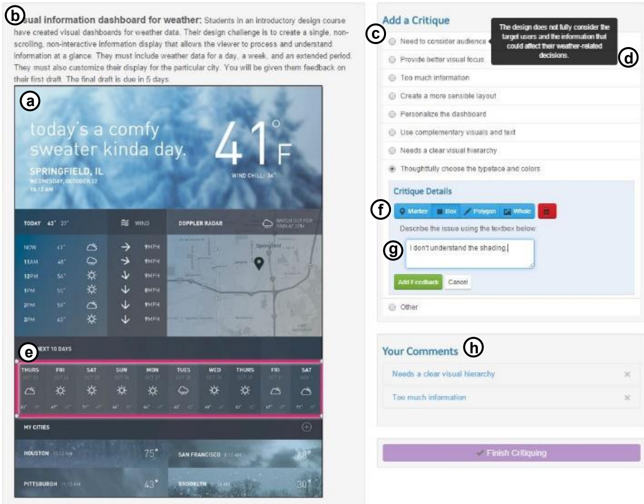

Figure 1. The feedback interface with rubric provided. See Apparatus for a description of the components.

METHOD

Apparatus

We used the CrowdCrit system [30] 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 includes (a) an image of the student design and (b) a paragraph describing the purpose of the design and experience of the designer. Workers produce critiques through the critiquing interface by first selecting (c) a relevant design principle from the rubric. Workers can view (d) descriptions for each principle by hovering over the design principle name. The selected principle forms the basis of the critique they wish to create. They can optionally provide (e) an annotation using (f) the toolbar to make a visual indication on the design. Additionally, they can provide (g) free-form comments to supplement and elaborate the pre-authored critique statements. Finally, workers can review (h) a list of their critiques 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.

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.

To generate critiques, we recruited 36 crowd workers of varying design experience, 12 from Upwork [43] and 24 from Amazon Mechanical Turk (MTurk). To help normalize the population's language skill, we restricted both pools of workers to consist of US-based workers only. Workers were then randomly assigned to critique either with or without the aid of a rubric. Upwork workers are typically more skilled and work on longer tasks than MTurk workers, so we had them critique 8 designs each and compensated them with $30. MTurk workers critiqued 4 designs each (half of Upwork) and were compensated $3 to match the expected pay rate of US minimum wage. These numbers ensured that each design received feedback from at least 3 workers in each pool and condition. On average, Upwork workers provided 4.3 critiques per design, and MTurk workers provided 2.0 critiques per design. On average, each design received 41 distinct critiques.

We carefully considered how much to pay participants, given that Upwork and MTurk offer different payment models and market rates. We could have matched hourly wages and offered MTurk worker exorbitant rates (or Upwork workers low rates), but this would have yielded rates that do not align with the market and would have introduced an additional confounding variable. For example, paying $10 for a task that normally pays $1 on a platform could have attracted particular types of workers, e.g., constantly underperforming workers, skewing our results [28, 32]. Further, by paying market

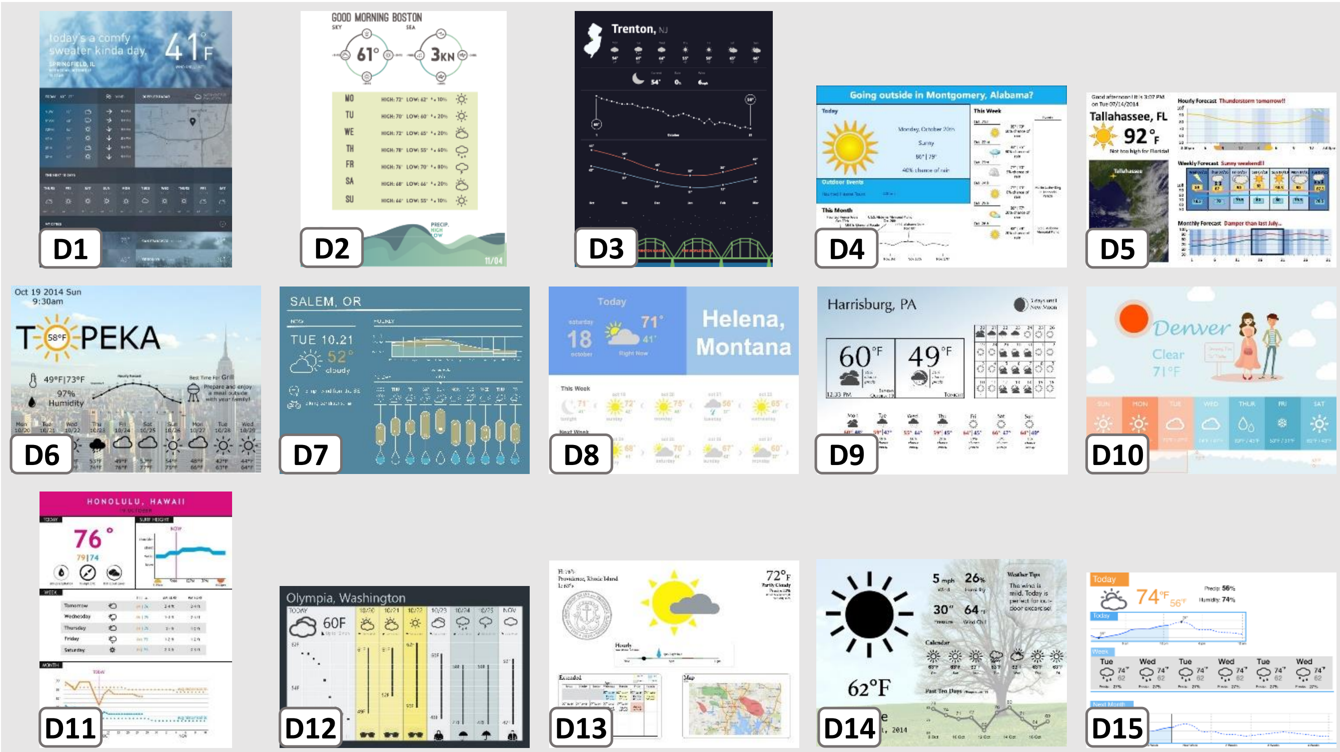

Figure 2. All 15 designs used in the experiment.

value on each platform, the study pragmatically compares the two platforms as designers would use them in the wild.

To determine expertise, we asked all workers to fill out a questionnaire on their previous design experience, including their design training and work experience. We define experts (12 in total) as workers with both a university degree and work experience in a design field; other workers are referred to as novices (24 in total). Under this definition, 11 out of 12 Upwork workers were experts and only one of 24 MTurk workers was considered an expert. Most MTurk workers (17) had neither a design-related degree or previous work experience in design.

The course instructor worked with our research team to create the rubric of design principles. See Table 1 for the full list of principles and descriptions. The principles were tailored to the assignment, and closely matched the official 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 helpfulness on a 1–10 Likert scale (10=best) for each point of feedback. After rating all critiques, students also optionally provided free-form opinions about the critiques.

Measures

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

Independent Variables

The first factor is worker *expertise* with two levels, *expert* and *novice*. Expert workers have both a design degree and prior experience 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.

Covariants

We control for two covariants. The student raters had different levels of design experience, which could have an impact on how they perceive the value of feedback. To operationalize design experience, we included a variable for the final course grades, ranging from 1 (lowest) to 4 (best). On the worker side, we likely recruited feedback providers with a wide range of English skills. Although we only allowed workers from within the US to take part in the experiment, we created a measure for vocabulary richness to control for this possible confound. To calculate vocabulary richness we removed all stop words and words not in *wordnet* from the critiques and drew random samples of 50 words from each feedback provider. We lemmatized all words using *NLTK* [2] and counted all unique lemmas. We then calculated the ratio of unique lemmas in these 50 word samples.

Dependent Variable

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 an interval scaled for

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 w/ rubric to D12 (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 w/ rubric to D12 (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 w/ rubric to D4 (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 w/ rubric to D13 (8)
This is not clear. – Novice w/ no rubric to D15 (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 w/ no rubric to D3 (9)
overall this is a great layout. – Expert w/ no rubric to D1 (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 w/ no rubric to D15 (10)

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

the purpose of analysis. Table 2 shows a sample of low and high-rated critiques.

RESULTS

To analyze main and interaction effects of rubrics and worker expertise on student ratings, we conducted an ANCOVA between our two factors: expertise (novice, expert) and rubrics (no rubrics, rubrics) with final students grades and vocabulary richness as covariates. In accordance with Harwell [19] and Schmider [37], we assumed our sample size n=34 and our substantial effect sizes (Cohens's d>0.6) to be sufficient to meet ANCOVA's normality criterion.

To ensure equal variance we conducted a Levene's test for homogeneity of variance, $F(5, 33) = 1.07$, $p = 0.39$, and it did not violate the equal variance assumption. Interactions between the covariate and the two independent variables expertise and condition were not significant $F(1, 33) = 1.46$, $p = 0.15$, which means that we can assume to have met the ANCOVA assumption of homogeneous regression slopes. We use Tukey's HSD test as our post-hoc method. The ANCOVA model requires us to adjust sub-population means for post-hoc testing. Table 4 reports the adjusted means and standard errors.

Presence of Rubrics Increase Critique Ratings

The ANCOVA results in Table 3 indicate that rubrics had a positive effect on rating. This finding is consistent with the results of the follow up Tukey HSD test as shown in Table 5.

Variable	SS	Df	F	p	sig.
(Intercept)	35.88	1	51.49	0.001	***
(C)ondition	4.14	1	5.95	0.02	*
(E)xpert	3.81	1	5.47	0.03	*
Grade	3.69	1	5.29	0.03	*
Vocabulary	0.06	1	0.12	0.73	
ExC	0.94	1	1.35	0.25	
Residuals	22.30	29			

Table 3. ANCOVA results of the main and interaction effects of Rubrics and Expertise on perceived helpfulness of feedback. Both independent variables are factors with two levels. Grade and Vocabulary are the covariants. * indicates significance ($p<0.05$) and * indicates significance ($p<0.001$).**

Rubrics	Expertise	M	SD	Adj. M	SE	low	high
no rubric	novice	5.74	1.28	5.76	0.25	5.25	6.27
	expert	6.83	0.41	6.79	0.25	6.10	7.49
rubric	novice	6.65	0.65	6.69	0.25	6.20	7.12
	expert	7.02	0.79	7.00	0.25	6.31	7.70

Table 4. Adjusted means calculated using the fitted ANCOVA model. The values for novices slightly increase while means for experts slightly decrease when correcting the model for the influence of students' final grades and workers' vocabulary richness.

		delta	p	low	high
rubric exp.	rubric nov.	0.38	0.78	-0.72	1.49
rubric exp.	no rubric exp.	0.21	0.97	-1.10	1.52
rubric exp.	no rubric nov.	1.28	0.02	0.13	2.43
rubric nov.	no rubric nov.	0.89	0.04	0.02	1.81
no rubric exp.	rubric nov.	0.17	0.97	-0.93	1.28
no rubric exp.	no rubric nov.	1.07	0.03	-0.08	2.22

Table 5. Tukey HSD results. The two left most columns describe the compared conditions. We abbreviate expert with exp. and novice with nov. The two right most columns indicate lower and upper bounds of the 95% confidence interval.

Experts Provide Better Critiques than Novices

As expected experts give feedback that is perceived as more useful than feedback from novices. Again both the ANCOVA (Table 3) and Tukey (Table 5) support our initial hypotheses.

Student Experience Influences Critique Ratings

As seen in Table 3, the experience of the student designer influences his or her rating of critiques. Students with very high final grades tend to give lower ratings than those with low final grades.

With Rubrics, Novices Do Not Differ from Experts

When experts and novices both use rubrics we do not find a significant difference between the groups (see Table 5).

Rubrics Help Novices More than Experts

We found that novices achieved significantly higher mean ratings with rubrics than without as shown in Table 5. Rubrics increased the average rating of reviews written by novices by 13.5%. Experts did not benefit from having rubrics as much as novices; we did not find a significant increase in ratings for experts with rubrics compared to experts with no rubrics.

Highly Rated Feedback Correlates with Linguistic Features

The first analysis indicated that rubrics had a positive effect on ratings of feedback written by novices. We wanted to understand what rubrics provide that lead to these results. To explore this, we conducted a linguistic analysis with a feature set previously used to investigate writing styles in an educational setting [23, 26]. We used the following subset of features: critique length (average word length, average sentence length), emotional content (valence and arousal), language specificity, and the grammatical mood of sentences.

We preprocessed all critiques with the NLTK part-of-speech (POS) tagger [2]. We then filtered stop words and words not in Wordnet [15]. Wordnet is a natural language tool that provides linguistic information on more than 170,000 words in the English language. We also lemmatized the remaining words to account for different inflections.

To estimate the correlation between our linguistic features, the presence of rubrics, and the observed ratings we use Pearson's product-moment correlation. We calculate ρ between each feature and our dependent variable (rating) and the independent variable (presence of rubrics). The features and results are described below.

Longer Sentences Receive Higher Ratings

The first two features we examined were the mean number of letters per word and mean number of words per sentence. For the mean word length, we considered only those words that have a Wordnet entry and are not stop words. The sentence length was measured including all words returned by the POS-tagger. All features positively correlated with higher ratings ($r(34) = 0.43$, $p<0.01$, $r(34) = 0.49$, $p<0.01$). We also found that critiques from the rubric condition had 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, $t(34) = 6.8$, $p <0.001$, $d = 2.24$ and $t(30) = 6.01$, $p <0.001$, $d = 2.02$.

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 the strength of valence. 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. We used *pattern.en*, a tool based on *NLTK*, to extract valence and arousal.

- 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 had 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 the specificity of words in the critique. 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 (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 ($M = 0.47$, $SD = 0.11$), $t(25) = 5.06$, $p<0.001$, $d = 1.74$.

Critiques that Question or Suggest Receive Higher Ratings

The last feature we considered is the grammatical mood 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. We again used *pattern.en* to extract sentence mood. Examples include:

- 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 ($M = 0.38$, $SD = 0.27$), $t(30) = 3.56$, $p<0.001$, $d = 1.20$.

The average activeness of a reviewer may sometimes not be as important as the total amount of actionable items. Therefore, we also measured the total amount of actionable items proposed in a review. We indeed found a correlation between number of action items (operationalized as total number of active sentences) and critique rating $r(34) = 0.514$, $p = 0.001$.

Language Differences between Upwork and MTurk

Our experiment recruited critique providers from from two different populations: MTurk workers and Upwork experts. We in fact found that almost all experts in our experiment were recruited through Upwork (11 experts, 1 novice) and almost all novices through MTurk (1 expert, 23 novices). The Cohen's κ for this correlation is almost perfect with $\kappa = 0.87$.

These marketplaces have different populations, most likely with differing commands of the English language. To account for this possibly confounding variable, we used vocabulary richness as a covariate. Furthermore we compared average vocabulary richness of both populations. We found no significant difference in average vocabulary richness between workers from Upwork (M = 0.34, SD = 0.07) and MTurk (M = 0.34, SD = 0.04) in our experiment $T(35) = 0.42$, $p = 0.67$.

Expertise does not Correlate with our Language Model

We also examined the correlation between the features and expertise of the worker. We did not find significant correlations between our language model and expertise. As our model can only explain certain dimensions of perceived helpfulness, we wanted to better understand what sets expert feedback apart in terms of content.

To this end, we examined and compared the highest rated feedback from experts with no rubrics and from novices with rubrics. We chose this subset of the feedback since it would provide the clearest distinction between how experts without rubrics and novices with rubrics produce helpful feedback. We coded all critiques rated 9 or 10 from these groups (37 expert and 15 novice critiques) 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, consider the highly rated feedback from an expert with no rubric and a novice with rubric in Table 2.

The expert feedback provider explained 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 also provides 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 found that roughly half featured a strong justification. Among the novice feedback, we found only about 20% featured a strong justification, though about 67% featured a weak justification. Sometimes the selected principle from the rubric acted as a justification, though in these cases it was more often a weak justification. These justifications partially account for why expert feedback is longer, and may also help explain why expert feedback is rated highly.

Qualitative Insights by Student Designers

After rating all comments, the participants answered an open-ended question about qualities they used to assess the helpfulness of feedback. In line with the linguistic analysis, many students appreciated feedback that made concrete suggestions. For example, participant D4 said "the comments that were most insightful were those which made concrete suggestions or examples of what I can do to improve my design." Conversely, feedback that critiqued the design without such concrete suggestions was judged to be unhelpful. For example, D12 disliked that "there were quite a few comments that just pointed things out that were good or bad (some very harsh), but no explanation as to how to improve."

While students in aggregate rated positive messages as helpful, some participants pointed out that positive messages may also serve a different role: they contribute towards a receptive disposition towards feedback, without being directly actionable. D1 wrote "while I enjoyed seeing the positive comments, it was tough to rate them on a scale of helpfulness". D11 reported "it was fun to get positive comments, but they weren't helpful at all. Makes me feel good but there's not much I can do with "clear layout."

The student designers also mentioned that repeated, consistent suggestions from multiple providers enabled them to prioritize issues. As D14 commented, "I found the feedback very useful in that I found emerging issues with my design that were noticed with multiple comments." D13 said, "I encountered a lot of repeated comments, which seemed a bit tedious to go through, but actually ended up just telling me what the most important parts I need to focus on are." In many crowdsourcing tasks, such redundancy may be viewed as wasteful or sub-optimal, but here the repetition helped designers focus on the areas that needed most attention.

DISCUSSION

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

RQ 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 significantly help the experts produce more valuable feedback for students. One potential explanation for this is that experts can already recall and apply design principles. They 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 trained 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 rubrics, but novices perform much better when they are 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 helpful feedback.

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

RQ 2: Designers Value Writing Style in Feedback and Rubrics Improve Writing Style

The latter half of the analysis looked at language features of the writing style in the crowd feedback text, and found multiple features that positively correlated with ratings. When we considered all possible combinations of the features, we found that the combination of arousal, valence, and specificity in particular achieved the highest correlation with rating. This correlational data suggest that the proper application of these features can produce higher value feedback. 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 hypothesis 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.

As hypothesized, we also found that positive valence correlated with higher ratings. This may be evidence of the conventional wisdom that it is better to point out both positives and negatives rather than being overly critical. Positive remarks can also be encouraging to the recipient [17, 51] and may be considered helpful 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 [36]. 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 or her comments to the particular design and designer. It seems reasonable that these qualities would improve the perceived helpfulness of the feedback.

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 did 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 for improving feedback by focusing on style rather than content.

Justification Matters

An unexpected result was that expertise did not correlate with any of the linguistic features in our analysis. 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) 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, would 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. 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 tradeoff here is that the more generally applicable a principle is, the less specific and precise it is for 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.

LIMITATIONS AND FUTURE WORK

Revisit Effects on Design Iteration

This study investigated the effect of rubrics on perceived helpfulness. Because our study took place in the naturalistic setting of an actual classroom assignment, where the student designers were also exposed to feedback from peers and instructors, we could not reliably measure the effects on the final design outcomes. Some studies [30, 50] have attempted to measure the effects of feedback on design outcomes with mixed results, but it remains a challenge to understand how rubrics ultimately affect design quality.

Further Explore Linguistic Analysis Findings

Our initial work on the linguistic analysis of feedback opens up a few avenues to explore. Our analysis provided correlational data, so the question remains as to whether these features have a causal relationship with perceived helpfulness. Another avenue would be to explore 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 for the provider to improve their critique. 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 or 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. In fact, Nguyen et al. [33] have already successfully applied a similar idea to help students localize their comments in peer reviews.

Further Analyze Expert Feedback

The linguistic analysis suggests how rubrics might add value to feedback but did not fully explain how experts produce valuable design feedback. Some initial qualitative analysis suggests that experts add clear and meaningful justifications to their critiques, leading to more cohesive pieces of feedback. Further investigation of 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 the Design Space of Structured Feedback

Our research corroborates the helpfulness of rubrics for novices. The particular rubrics employed in our study were provided by course instructors and matched the particular design assignment. Luther's prior work used more general rubrics derived from instructional texts about visual design [30]. Both are guided by Sadler's requirements for effective formative feedback [36]. However, a larger design space of rubrics in particular, and of ways to structure feedback more generally, exists. A natural follow-up would be to investigate different strategies for structuring feedback tasks and their trade-offs. This can deepen our understanding of the role of rubrics and other task structuring techniques in crowd feedback systems.

CONCLUSION

Crowd feedback systems that recruit novices to provide feedback have the potential to affect a wide range of designers, but existing research had yet to evaluate their value compared to hiring experts. Our experiment provides evidence that arming novice feedback providers with an expert rubric helps them produce feedback of similar value as expert feedback.

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. Experts did not necessarily produce feedback with better writing style, but they provided stronger and clearer justifications for their critique. These findings motivate further investigation into how feedback systems can structure high-quality feedback.

ACKNOWLEDGEMENTS

The authors thank Karen Berntsen for her valuable design expertise in authoring the rubrics, as well as the students and crowd workers who participated in our study. Financial support provided by the National Science Foundation under IIS grants 1210836, 1208382, 1217096, and 1122206, and by the German Academic Exchange Service through the FIT Worldwide program.

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. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2399–2402. DOI: http://dx.doi.org/10.1145/1753326.1753688

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

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

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

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

16. Gerhard Fischer, Kumiyo Nakakoji, Jonathan Ostwald, Gerry Stahl, and Tamara Sumner. 1993. Embedding Computer-based Critics in the Contexts of Design. In *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems (INTERCHI '93)*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 157–164. http://dl.acm.org/citation.cfm?id=164632.164891

17. 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

18. 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 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 235–244. DOI: http://dx.doi.org/10.1145/2757226.2757249

19. 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

20. 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

21. 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

22. 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

23. 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

24. Joy Kim, Mira Dontcheva, Wilmot Li, Michael S. Bernstein, and Daniela Steinsapir. 2015. Motif: Supporting Novie Creativity through Expert Patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 1211–1220. DOI: http://dx.doi.org/10.1145/2702123.2702507

25. 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

26. 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

27. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20, 6 (Dec. 2013), 33:1–33:31. DOI: http://dx.doi.org/10.1145/2505057

28. John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*. 17–20. http://ir.ischool.utexas.edu/cse2010/materials/leetal.pdf

29. 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.

30. 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

31. 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

32. Winter Mason and Duncan J. Watts. 2010. Financial incentives and the "performance of crowds". *ACM SIGKDD Explorations Newsletter* 11, 2 (May 2010), 100. DOI: http://dx.doi.org/10.1145/1809400.1809422

33. Huy Nguyen, Wenting Xiong, and Diane Litman. 2014. Classroom Evaluation of a Scaffolding Intervention for Improving Peer Review Localization. In *Intelligent Tutoring Systems*, Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia (Eds.). Number 8474 in Lecture Notes in Computer Science. Springer International Publishing, 272–282. http://link.springer.com/chapter/10.1007/978-3-319-07221-0_34

34. David A. Robb, Stefano Padilla, Britta Kalkreuter, and Mike J. Chantler. 2015. Crowdsourced Feedback With Imagery Rather Than Text. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 1355–1364. DOI: http://dx.doi.org/10.1145/2702123.2702470

35. 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.

36. 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

37. 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

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

39. 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.

40. Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 275–284. DOI: http://dx.doi.org/10.1145/1958824.1958865

41. 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

42. 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

43. Upwork. Upwork, the world's largest online workplace. (2015). https://www.upwork.com

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

45. 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

46. 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.

47. 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

48. 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

49. 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

50. 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

51. Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 407–416. DOI:http://dx.doi.org/10.1145/2145204.2145269

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