Structuring, Aggregating, and Evaluating Crowdsourced Design Critique

Kurt Luther[1], **Jari-Lee Tolentino**[2], **Wei Wu**[3], **Amy Pavel**[3], **Brian P. Bailey**[4], **Maneesh Agrawala**[3], **Björn Hartmann**[3], **Steven P. Dow**[1]

[1]Carnegie Mellon University {kluther,spdow}@cs.cmu.edu

[2]University of California, Irvine jltolent@gmail.com

[3]University of California, Berkeley {amypavel,maneesh,bjoern}@eecs.berkeley.edu

[4]University of Illinois at Urbana–Champaign bpbailey@illinois.edu

ABSTRACT

Feedback is an important component of the design process, but gaining access to high-quality critique outside a classroom or firm is challenging. We present CrowdCrit, a web-based system that allows designers to receive design critiques from non-expert crowd workers. We evaluated CrowdCrit in three studies focusing on the designer's experience and benefits of the critiques. In the first study, we compared crowd and expert critiques and found evidence that aggregated crowd critique approaches expert critique. In a second study, we found that designers who got crowd feedback perceived that it improved their design process. The third study showed that designers were enthusiastic about crowd critiques and used them to change their designs. We conclude with implications for the design of crowd feedback services.

Author Keywords

Design; critique; feedback; social computing; crowdsourcing

ACM Classification Keywords

H.5.3 Group and Organizational Interfaces: Computer supported cooperative work, web-based interaction

INTRODUCTION

For centuries, critique has provided a foundational exercise for art and design, and also more recently, for project-based education in computing and engineering [23]. The traditional studio critique is a co-located communication activity where someone presents preliminary work and then critics—often teachers and peers—provide feedback to improve the design [5, 2]. Critiques can help novices to understand key principles in a domain [12], to compare alternatives [7, 27], to articulate the goals and assumptions behind their work, and to recognize how others perceive their work [15]. The critique providers also learn by developing domain-specific vocabulary [4] and rehearsing the mechanics of the crit process.

CSCW 2015, March 14–18, 2015, Vancouver, BC, Canada.

ACM 978-1-4503-2922-4/15/03 ...$15.00
http://dx.doi.org/10.1145/2675133.2675283

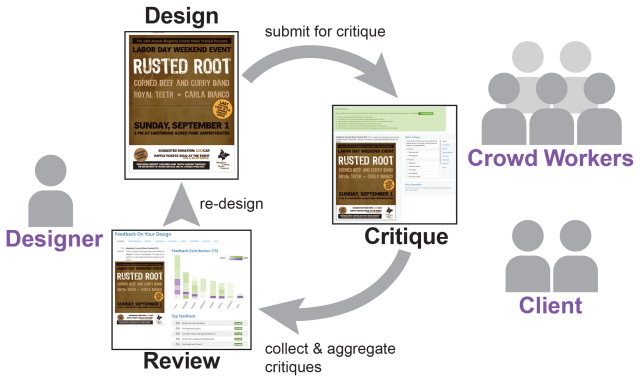


Figure 1. CrowdCrit allows designers to submit preliminary designs to be critiqued by crowds and clients. The system then aggregates and visualizes the critiques for designers.

Critique leads to knowledge sharing and helps inculcate important values and aesthetics.

Feedback can be valuable to a wide audience, from professional designers to design students to everyday designers working on slide decks and flyers. Unfortunately, high-quality critique can be difficult to obtain outside of a design firm or classroom. There are numerous online design communities where people seek feedback (e.g. Forrst [14], Photosig [32]), but these often yield sparse, superficial comments [32]. Some (e.g. Dribbble [18]) are invitation only, and most require members to build a reputation and feel comfortable sharing preliminary work. Novice designers in particular may experience evaluation apprehension and avoid sharing their works-in-progress alongside experts [18].

Recently, a number of academic and commercial efforts have explored how to leverage paid crowdsourcing platforms like Amazon Mechanical Turk (MTurk) within a design process. These platforms are attractive because they provide fast, scalable access to human intelligence at a reasonable cost, but most crowd workers lack design knowledge and cannot provide useful critiques. Most of these efforts, therefore, focus on using crowds to provide high-level impressions [11, 13] or learn more about a target audience [8]. Others (e.g. Voyant [33]) decompose the feedback process into microtasks but do not focus on the process and language of design critique.

In this paper, we propose and evaluate a different approach. We built a web-based system called CrowdCrit (Figure 1) that allows designers to submit preliminary designs, receive critiques from many crowd workers, and explore the aggregated results using an interactive visualization. Rather than searching a large crowd for design experts, CrowdCrit provides a structured interface, inspired by scaffolding theory, to help non-expert crowds submit critiques. CrowdCrit integrates with MTurk to leverage the speed, scale, and cost benefits of paid crowds, but is designed to work with any type of crowd (classmates, coworkers, user communities, etc.).

We evaluated CrowdCrit with three studies, focusing on the experience of designers and the characteristics of crowd critiques. The first study compared the characteristics of crowd critiques (N=14) to expert critiques (N=3) of the same set of poster designs. Crowd and expert critiques had comparable internal consistency, though it was low. The results also show that aggregating crowd critiques trends towards approximating expert critiques, with 10–14 crowd workers finding 40–60% of design issues identified by experts.

The second study (N=14) used interviews to explore designers' reactions to crowd critiques and examples and patterns of use in the context of a poster design contest. We found that designers of varying experience levels considered crowd critiques valuable and used them to make changes to their designs. We describe how designers explored and interpret crowd critiques, weighing attributes like crowd expertise and critique frequency against their own intuitions.

The third study (N=18) is, to our knowledge, the first controlled experiment to examine the specific impact of crowd-generated design feedback against a baseline. In a second contest, half of designers received crowd critiques and half received generic feedback. The study found that with crowd critiques, designers reported noticing more issues, making more changes, and producing an overall better design. However, third-party ratings of designs found few conditional differences, suggesting that CrowdCrit may lead designers to focus on refinement, rather than broader changes.

In the following sections, we review related work on crowdsourcing and feedback and describe the CrowdCrit system and our evaluations. We conclude with design implications for crowd feedback services.

RELATED WORK

Theories of learning and assessment

To inform our interface design, we draw on learning theory literature that explores how best to teach new concepts to students. *Scaffolding* refers to a teaching technique where learners are given significant support — by teachers or computer-based tutors, for example — to help them learn new material [25, 3]. Our system, inspired by scaffolding theory, structures the visual design critique process for the crowd by suggesting design issues and concepts as a set of pre-authored critique statements.

Sadler reviewed different feedback approaches and argued that good feedback must help a student grasp the concept of a standard (*conceptual*), compare the actual level of performance with this standard (*specific*), and engage in action that closes this gap (*actionable*) [22]. Our system structures crowd workers to include all three components in the context of visual design critique.

Peer critique vs. external feedback

Recently, researchers have explored the viability of online critique within an educational context. Tinapple et al. [26] developed CritViz, a system which enables peer critiques of designs. Kulkarni and Klemmer [16] also developed a peer critique system for use in a Massive Open Online Course (MOOC), and Easterday et al. [10] developed a mixed face-to-face and online critique system for design students in a classroom environment. These systems were shown to be effective in a formal educational setting, with students of fairly uniform levels of knowledge and motivation. Alternatively, Dow et al. [8] explored the utility of external crowds to contribute outside perspectives to add authenticity to a design course. CrowdCrit seeks to leverage the diverse capabilities of crowd workers and focuses their efforts on producing visual design critique.

Reputation vs. anonymity

Many commercial sites like Dribbble [18] and Forrst [14] allow members to upload their own creative projects and provide feedback on each other's work. However, novices are likely to experience evaluation apprehension and feel intimidated sharing their work alongside professionals [18]. CrowdCrit operates in a double-blind fashion, so that designers do not feel reluctant to share and critique providers are not influenced by the designer's reputation.

Payment vs. reciprocity

Systems that follow the reciprocity model do not require a financial transaction, but they often fail to produce enough useful feedback in a timely manner. Xu and Bailey [32] studied critique behaviors in the online photography community PhotoSIG, where members voluntarily critique each other's photos and hope for reciprocity. They found that 80% of photos received less than four critiques, which most users found insufficient. The paid crowdsourcing paradigm used by CrowdCrit provides a novel opportunity to deliver large quantities of feedback quickly and at a reasonable cost.

Impressions vs. structured feedback

Sites like Five Second Test [13] and Feedback Army [11] allow users to pay crowds to provide feedback on their designs or websites, typically in the form of overall impressions or general reactions. This kind of feedback is valuable for getting an audience's perspective on a design, but it does not illuminate why a design works well, or suggest what to change. Even in online communities for professional designers, shallow reactions, rather than detailed critique, are the norm [32].

Xu et al. [33] created the Voyant system to obtain crowd feedback on visual designs by creating structured micro-tasks. Most of the feedback types supported by Voyant centered around audience reactions (e.g. which elements are noticed first, general impressions of the design). The system also paid

the crowd to rate, on a Likert scale, how well the design meets the guidelines of alignment, contrast, proximity, and repetition. CrowdCrit builds on these ideas by structuring not only the crowd's ability to *recognize* a wide range of design issues, but also the *process* and *language* of constructing a critique.

Aggregating crowd feedback

Most commercial and academic critique software is not designed to scale to large quantities of critique. CritViz [26] supported peer critique in a college course of 75 students, but each assignment received only five critiques. Voyant [33] supports large quantities of crowd feedback, using visualizations such as word clouds and histograms. Our research builds on this idea by using a stacked bar chart visualization to aggregate rich critique data, including valence (positive/negative), text comments and graphical annotations, and expertise across seven design principles.

General solutions for aggregating crowd-generated content are also germane. For example, CommentSpace [30] provided interactive visualizations and structure to help crowds perform social data analysis. We implemented several of these papers' recommendations, such as linked discussions and structured comments to support deeper analysis. Our system differs in that it focuses on the unique requirements of design critique and operates as a service for designers, providing distinct interfaces for critique collection (from crowds) and critique aggregation (for designers).

PRELIMINARY STUDIES

To better understand the challenges of crowd-generated critique and the appropriate type of structure, we conducted a series of preliminary studies on MTurk (Figure 2) to see what type of feedback crowds provided, given minimal training or structure. In our first pilot, we paid three crowd workers to write open-ended feedback on a set of eight designs. A quick analysis showed that only five of 65 responses fulfilled Sadler's [22] requirements for high-quality feedback. Next, we added some structure to the task, asking three different workers to provide feedback on four designs using two text boxes: one for describing the problem and one for explaining why it was a problem. However, these workers still failed to identify and justify significant issues with the designs. In our third pilot, we embedded design knowledge into the structure, asking three new workers to critique nine designs by considering a list of pre-authored critique statements and checking the box next to ones that applied to the given design. This approach showed promise; the three workers agreed on 65% and 52% of the issues in the two checklists we provided.

Overall, these pilots suggested that eliciting open-ended critique without appropriate task structuring led to predominantly low-quality responses that lacked conceptual grounding and action-oriented advice. Introducing structured checklists was more successful and suggested that given this type of domain-specific structure, crowd workers might identify legitimate design problems and generate reasonably useful feedback. These early findings, along with the prior work reviewed above, informed the design of our CrowdCrit system, described in the next section.

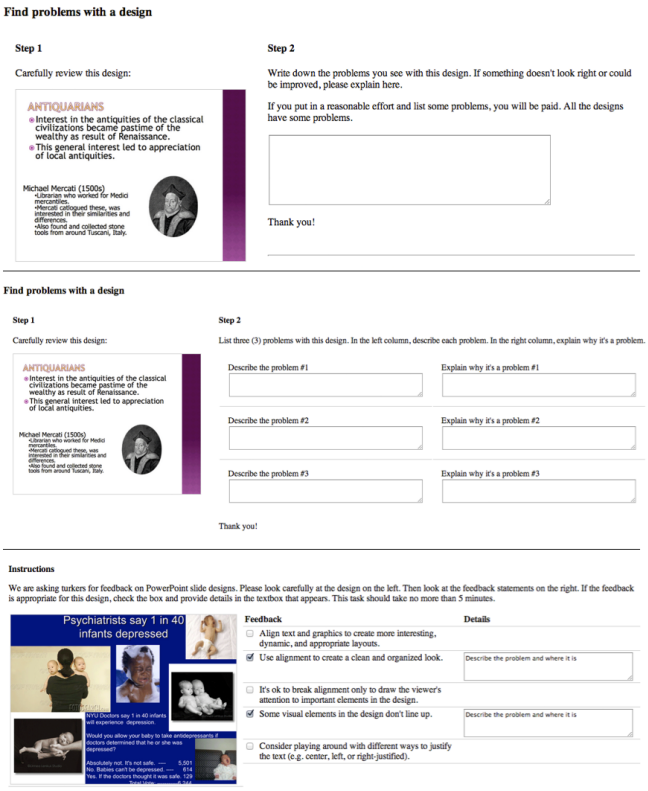


Figure 2. Crowdsourcing interfaces used for the first three pilot studies that indicated some drawbacks of open-ended designs feedback and benefits of additional structure.

THE CROWDCRIT SYSTEM

CrowdCrit is a web-based system built in Python and JavaScript that allows a user to request, receive, and review design critiques generated by an online crowd. To use the system, a designer uploads an image and provide some contextual information, such as a title, brief description, and intended audience. Next, crowd workers provide feedback on the design using a scaffolded critique interface. Finally, the designer can explore the crowd critiques using an aggregation interface which includes an interactive visualization. This process can be repeated whenever the designer seeks a new round of feedback.

Eliciting critique

CrowdCrit is designed to elicit valuable critiques from any online crowd, even those with limited design experience. Our preliminary studies and prior work suggested a structured interface could help novice crowds adopt the process and language of visual design critique. This required solving two problems: first, generating structured design knowledge based on best practices; and second, embedding this material in an interface suitable for non-expert crowd workers.

Critique statements

There is no universally agreed-upon reference for visual design critique, but our survey of widely cited design textbooks and other resources revealed significant overlap in terminology, concepts, and best practices. Therefore, two authors

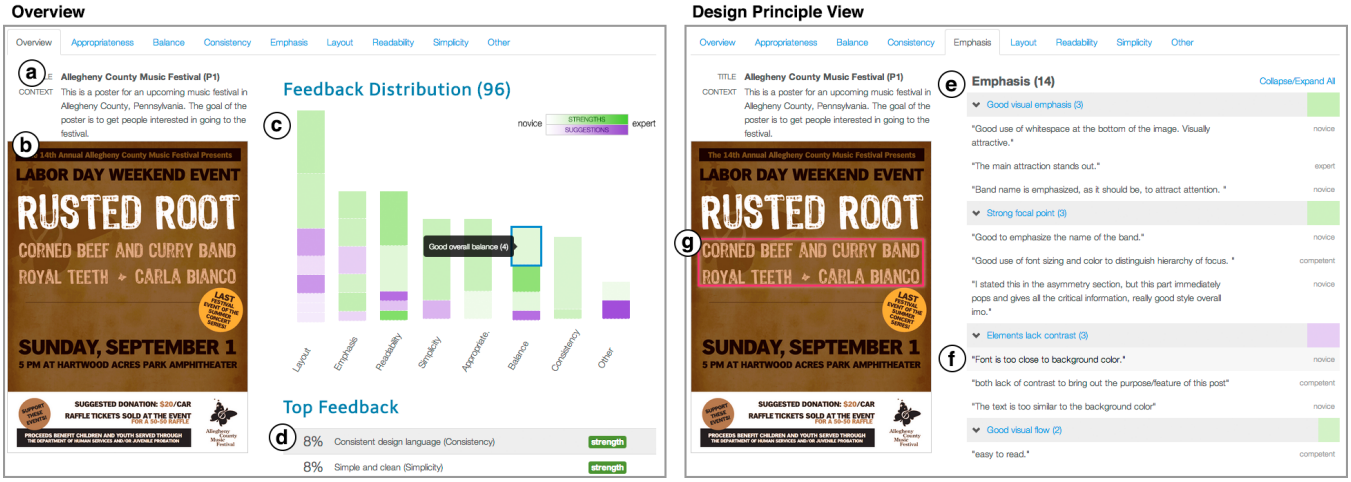


Figure 4. The Overview (left) and Design Principle view (right) for the CrowdCrit aggregation interface.

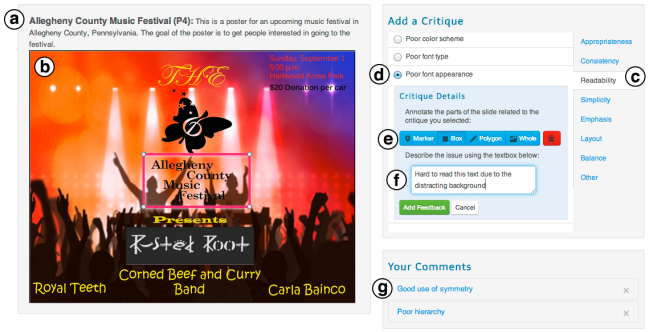


Figure 3. The CrowdCrit critique interface.

with design experience performed a bottom-up analysis of key sources [21, 31, 17, 6] to generate material for structuring our critique interface. They began by reviewing each text for lists of best practices or principles and extracting these along with definitions and examples. Next, they grouped related concepts together based on overlapping terminology, similar examples, their own knowledge and experience, and direct references in the text. After several iterations, they produced a list of seven key design principles: *Readability, Layout, Balance, Simplicity, Emphasis, Consistency,* and *Appropriateness*. The principles are distinct, but not mutually exclusive, as there are often multiple effective ways to identify issues with a design. This general approach is inspired by similar efforts in the HCI community to produce other types of design principles (e.g. [20, 28, 29]).

Each principle is composed of a set of pre-authored "critique statements" about more specific issues. CrowdCrit follows from traditional critique [12] and provides both positive and negative critique statements. For example, critique providers can praise the appropriateness of the design ("Reaches intended audience") or raise concerns about it ("Wrong audience"). A total of 70 critique statements are available across the seven principles.

Finally, each critique statement has a short name and a longer, more detailed description. The description identifies the specific issue, connects it to a broader design principle, and (in most cases) offers a generic solution. This format is meant to embody Sadler's [22] criteria for good feedback and provide a basic level of utility. For example, within the *Simplicity* principle, there is a critique statement with the name "Too much text" and the description: "The abundance of text makes this difficult for viewers to comprehend. Consider condensing this text by focusing on the essential message."

Critique interface

Crowd workers begin by completing a survey of self-reported design experience and a design knowledge quiz, used to categorize their expertise as low, moderate, or high. Next, workers are presented with the critique interface (Figure 3). The design's contextual information (a), including intended audience, along with the design itself (b), appears on 1 side, and the critique tools are shown on the other. The interface uses progressive disclosure [24] to avoid overwhelming novice critique providers with information about design. There are tabs for each of the seven design principles, plus an "Other" tab for critiques that our list may overlook (c). Crowd workers click a tab to reveal the critique statements for that principle and hide the others (d). Workers can mouseover the short statement names to display the longer description as a tooltip. After selecting a statement, the worker can use annotation tools, such as markers or polygons, to indicate the relevant area of the design (e). Workers can also elaborate on the critique using a text box (f). Finally, the worker saves the critique, and it appears on the comment list (g). The worker can repeat this process for as many critiques as they desire.

Aggregating and presenting critique

Once workers have critiqued the design, the designer needs to be able to review the feedback in a meaningful way. This poses interesting challenges compared to a traditional studio critique, which unfolds linearly, in real time, from a small number of co-located participants. In contrast, CrowdCrit

generates potentially large quantities (50+) of individual critiques from online, distributed providers in a nonlinear, asynchronous format. We needed to design an interface that would handle these requirements and allow the user to identify the most significant issues with a design.

Aggregation interface

The CrowdCrit aggregation is shown in Figure 4. Like the critique interface, it leverages progressive disclosure to avoid overwhelming the user with details. The design and contextual information appear on one side (a, b), and an interactive visualization providing an overview of all critiques as a stacked bar chart (c) appears on the right (Figure 4, left). Each bar represents a design principle, and each bar segment indicates a critique statement chosen by workers. Bars are ordered by decreasing critique count so that the principle with the most critiques appears first. Color encodes critique valence; positive critiques are green and negative critiques are purple. We use lightness values to encode crowd expertise; darker shades indicate critiques from workers with higher expertise. Below the visualization, a "Top Feedback" section lists the most frequently used critique statements (d).

The designer can click a bar or segment to drill down on the detailed critiques for that principle (e), ordered by decreasing frequency (Figure 4, right). The interface displays each critiquer's text comment and expertise (f); we label workers in the bottom quartile of design expertise as novices, workers in the middle quartiles as competent, and workers in the top quartile as experienced. Hovering over a critique causes any corresponding annotations to appear on the design itself (g).

STUDY 1: COMPARING CROWD VS. EXPERT CRITIQUE

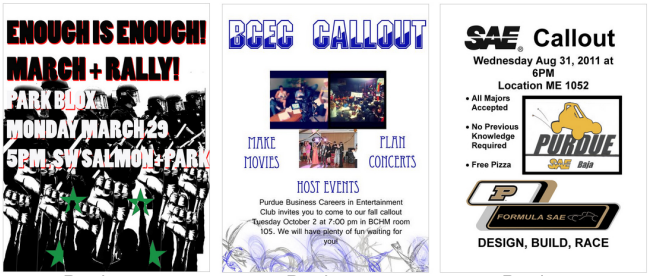
Our first study focuses on how crowd critiques compare to expert critiques. We consider the following research questions:

- How internally consistent are crowd vs. expert critiques?
- How similar are individual and aggregated crowd critiques to expert critiques?

Methods

We designed a controlled experiment using three event poster designs gathered online (Figure 5). Each poster received critiques from a group of 14 crowd workers and a group of three expert designers. The crowd workers came from MTurk and were recruited with the same criteria described above. The expert designers all had degrees in design and had worked full-time as professional designers. Both crowd workers and experts provided critiques independently, using the CrowdCrit interface.

The crowd produced between 38 and 58 critiques for each of the three designs (approximately four crits per design), using between 25 and 33 unique critique statements. Experts generated between 21 and 24 critiques for the designs, using 18–20 different statements. Finally, we generated a set of random critiques as another baseline. We generated 14 groups of four random critiques per design, to approximate the quantity of crowd-generated critiques.



Design 1 Design 2 Design 3

Figure 5. Poster designs for Study 1, used to advertise real events (collected from the web).

To measure consistency among more than two raters, we used Fleiss's κ. For each of the 70 possible critique statements offered by CrowdCrit, we calculated how many critiquers used that statement, and how many didn't.

To compare crowd and random critiques to experts, we use an approach inspired by Nielsen and Molich [20] to validate usability heuristics. For each design, we produce a "gold standard" list of expert critiques which is the union of the statements chosen by each expert independently. We chose the union rather than intersection due to the low levels of agreement observed among experts, described below. This yielded 18, 19, and 20 "correct" statements for designs 1, 2, and 3, respectively. If crowd or random critiques include the same statements as experts, we can think of this as a "match" or as both groups identifying the same issues and problems.

Finally, to understand the impact of false positive, we calculate precision, recall, and f-measure (combined precision and recall) for crowd and random critiques vs. expert critiques.

Results and discussion

Internal consistency of crowd vs. expert critiques

The crowd's κ scores ranged from 0.04 to 0.09, while experts ranged from -0.05 to 0.20. Random critique κ scores ranged from -0.01 to 0.01. All of these scores are considered poor or slight agreement. Yet, across all designs, crowd workers converged on a subset of statements ($\leq$33 of the available 70), and on average each statement that was used received two different crits. The highest crowd agreement occurred in design 1, where six of 14 workers (43%) chose the statement, "Lacks background contrast."

Individual crowd workers vs. experts

We found the average worker identified 6–8% of expert critiques, though workers provided only about four crits per design. The best workers produced crits that matched 17–25% of the expert list. Random critiques fare worse, with average matches ranging between 3–6%, and a maximum of 11%, even with equal or higher numbers of critiques vs. the crowd condition.

Aggregate crowd workers vs. experts

While individual crowd workers provide relatively few matching critiques, one of the benefits of crowdsourcing platforms is that we can easily recruit more workers and generate

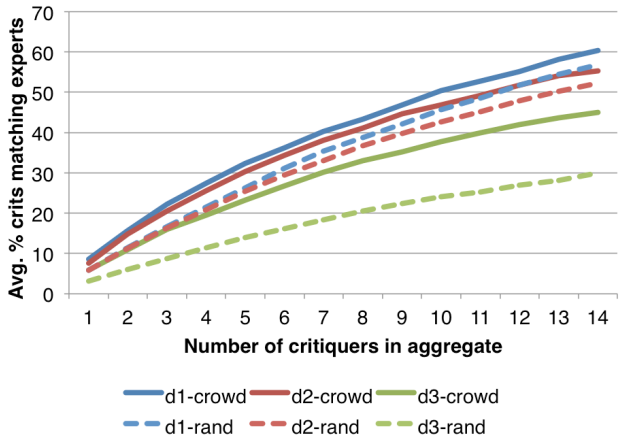


Figure 6. Proportion of expert critiques matched by aggregated crowd vs. random critiques.

more critiques. However, we first need to understand the relationship between more workers and better expert matches.

For design 1, the crowd identified 60.3% of the experts' selections, compared to 56.9% of random crits. For design 2, the crowd found 55.3% of the expert issues, compared to 52.2% for random crits. For design 3, the crowd found 45% of the issues, compared to 30% for random crits. These results suggest that more workers produce better results (8x to 10x more matches). Data on >14 workers is needed to determine if higher matches are possible.

Figure 6 shows how match percentage changes with aggregation. For each design, we ran 500 trials randomizing the order in which the crowd critiques were aggregated. Overall, the trend suggests a linear relationship between number of crowd critiquers and proportion of expert issues identified. An aggregate of 10 workers yields 40-50% of the issues, while 14 workers find 45–60% of the problems.

Precision/recall scores for crowd critiques

The above analyses consider how well crowds match the results produced by experts, but they don't address crowd critiques that don't match the experts, i.e., false positives. Figure 7 graphs the precision, recall, and f-measure scores for crowd and random crits across all designs. The crowd achieves better scores across the board, on average 0.12 points higher (out of 1.0) for precision and recall. This supports the above evidence that crowd critiques match experts better than random, and also suggests that crowds miss fewer expert crits and offer fewer false positives.

However, not all false positives are bad critiques. Nielsen and Molich [20] observed few false positives, and in fact revised their expert list post-hoc to include new issues identified by novices. Similarly, our informal analysis finds many crowd critiques that point out legitimate, if minor, issues.

Overall, compared to random critiques, crowd critiques show higher internal consistency, individual workers identify a greater percentage of expert issues, and aggregate percentages are also higher. Additionally, crowd critiques may provide richer feedback, in the form of optional annotations and text comments, that random critiques lack. The first study suggests that crowd critiques could benefit designers, but we know little about how designers might respond to crowd critiques or incorporate them into their actual design process. We address these issues in the following two studies.

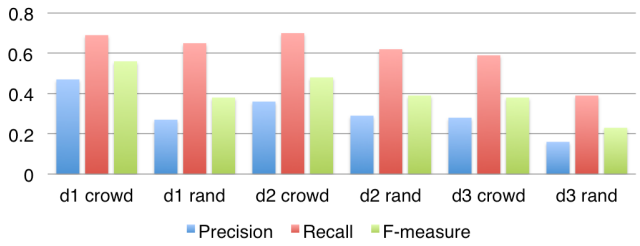


Figure 7. Precision, recall, and f-measures for crowd and random critiques vs. expert critiques.

STUDY 2: DESIGNER REACTIONS TO CROWD CRITIQUE

This exploratory study examines how designers leverage CrowdCrit in the context of a visual design contest. It seeks to answer three key research questions:

- How do participants react to the critiques and the aggregation interface?
- How do participants react to the source of crowd critiques?
- How do participants use crowd critiques?

To facilitate rich data collection and comparison across participants, we organized a poster design contest with real clients (organizers of a local music festival) and real prize money ($250 for the best poster). Participants used their own laptops and software and were provided with crowd feedback from CrowdCrit.

Method

We recruited 14 participants (six female) ranging in age from 20 to 33 ($\mu = 25$) by posting flyers at a university campus and advertising on neighborhood email lists, Craigslist, and Reddit. We sought to recruit participants with a wide range of experience, which we later categorized into three groups: *novice* designers, who had little or no design experience; *competent* designers who had taken a design course or had done some freelance design work; and *experienced* designers who had earned a degree in design or had worked full-time as a designer. To motivate participants while acknowledging their primary role as research participants, we compensated each participant $10/hour ($30 total), the standard rate for research participation at our institution, and awarded the $250 contest prize to the winning participant. The study consisted of two design sessions separated by a week to provide time to generate feedback from both the crowd and client.

Design session 1

In the first design session, each participant came to our lab and completed a pre-survey about their design experience. We then explained the goals of the contest and presented a design brief, written in collaboration with the clients, which included information about the event and other requirements,

such as using the festival logo. Participants then had one hour to design a draft of their posters.

Crowd and client critique

After all participants had completed the first design session, we used CrowdCrit to gather crowd and client critiques for each design.

To gather crowd critiques, we recruited 50 crowd workers on MTurk and paid each the US minimum wage ($7.25/hour, or $3.50 for this 25-minute task). All workers were US-based to reduce the impact of cultural differences on perceptions of design. After completing a pre-survey and a tutorial video, each worker critiqued three randomly selected posters from the group, completed a brief post-survey about their opinions of the system, and received payment.

To gather client critiques, we arranged an hour-long, face-to-face meeting with two of the festival organizers. They discussed each of the 14 initial designs and provided feedback using the same critique interface as the crowd. We also asked them to share their impressions of the critique and aggregation interfaces.

Design session 2

After one week, each participant returned to the lab for a second design session. We demonstrated the aggregation interface and asked participants to explore the crowd and client critiques using a think-aloud protocol. Participants then had another hour to iterate on their original version and submit a final design, using the aggregation interface and critiques as much or as little as they wanted. We then conducted a semi-structured interview covering design process, thoughts on the crowd and client critiques, and experience with the interface. The combination of think-aloud and interview lasted, on average, 27 minutes per participant. Pre- and post-critique versions of all 14 designs can be viewed in Figure 8.

Client judging

We set up a second face-to-face meeting with the clients where we showed them each participant's first draft, the client and crowd critiques, and the final design. We asked clients to comment on how the designs had changed and how well the participants had responded to their feedback. The clients also voted on a winning poster (P6).

All client and participant interview data was audio-recorded, fully transcribed, and analyzed using a bottom-up approach. In our first pass, we annotated quotes which addressed one of our research questions above, and in several subsequent passes, we grouped related quotes together to elicit broader themes, presented in the next section.

Results and discussion

How do participants react to the critiques and the aggregation interface?

Quality: On the whole, participants found the majority of critiques to be helpful, especially given the crowd's minimal qualifications and the lack of moderation. They mentioned specific reasons such as causing them to notice issues they hadn't considered (P5), appreciating the concreteness and rich detail (P6), and feeling like the critiques were carefully considered (P8). Some participants particularly appreciated positive critiques; P6 noted, "I tend to always second guess my work, so hearing affirmation, hearing people say 'This is a strength,' it helps a lot."

Quantity: Overall, each poster received an average of 69 ($\sigma = 12$) critiques, where 26 ($\sigma = 17$) critiques were positive and 43 ($\sigma = 20$) critiques were negative. Clients contributed 4% of all critiques. The majority of participants were satisfied by the number of critiques they received, with some expressing pleasant surprise at receiving more than they expected. P7 characterized the general attitude towards quantity, saying, "The more feedback, in my opinion, the better, because there's a chance that you'll be able to see a larger trend." P1 agreed, "It feels like if it's only two people [critiquing], that's just their taste. [With CrowdCrit] you get a sense of, 'Overall, people like the alignment,' [or] 'Overall, people like this.'"

Speed: We asked participants how quickly they'd prefer to receive crowd critiques, and most responses ranged between one day and several days for each round of feedback, depending on the project. For this study, the CrowdCrit system generated critiques for all 14 posters in less than two hours, or on average, less than 10 minutes per design. We shared this with participants, and most found the turnaround time to be surprisingly fast. In P1's words, "If you can do it in an hour, it's perfect—I think that will really justify the price."

Cost: The total cost to collect critiques for all 14 designs was $175, or $12.50 per design on average; workers were paid an hourly rate equivalent to minimum wage in our location. We asked participants how much they'd be willing to pay for the critiques of their poster drafts. For novices, who typically designed for themselves, the most common response was a flat rate of around $5–10 per round of feedback. Given the abundance of feedback provided in this study, we speculate that a service like CrowdCrit can provide value at the $5–10 price point.

Interface Participants generally found the aggregation interface easy to use. They described it using phrases such as "easy to understand" (P1), "simple and clean" (P5), "clear-cut" (P6), and "intuitive" (P7). P12 praised the "straightforward" design, saying, "the brilliance of this interface is the fact that it's really simple."

One of the most popular features was the ability to hover over individual critique text to reveal the corresponding graphical annotations (e.g. markers or selected regions) on the design itself. P6 liked that annotations added specificity to ambiguous comments; without them, critique statements like "Lacks balance" would seem, in her words, "fake."

How do participants react to the source of crowd critiques?

Crowd expertise: When exposed to critiques from a variety of sources, participants usually considered client critiques first and crowd critiques second, often commenting that client satisfaction trumped everything else. Among crowd critiques, participants paid closer attention to expert critiques because they perceived them to be higher quality. Responses to novice feedback were more varied. Some experienced designers

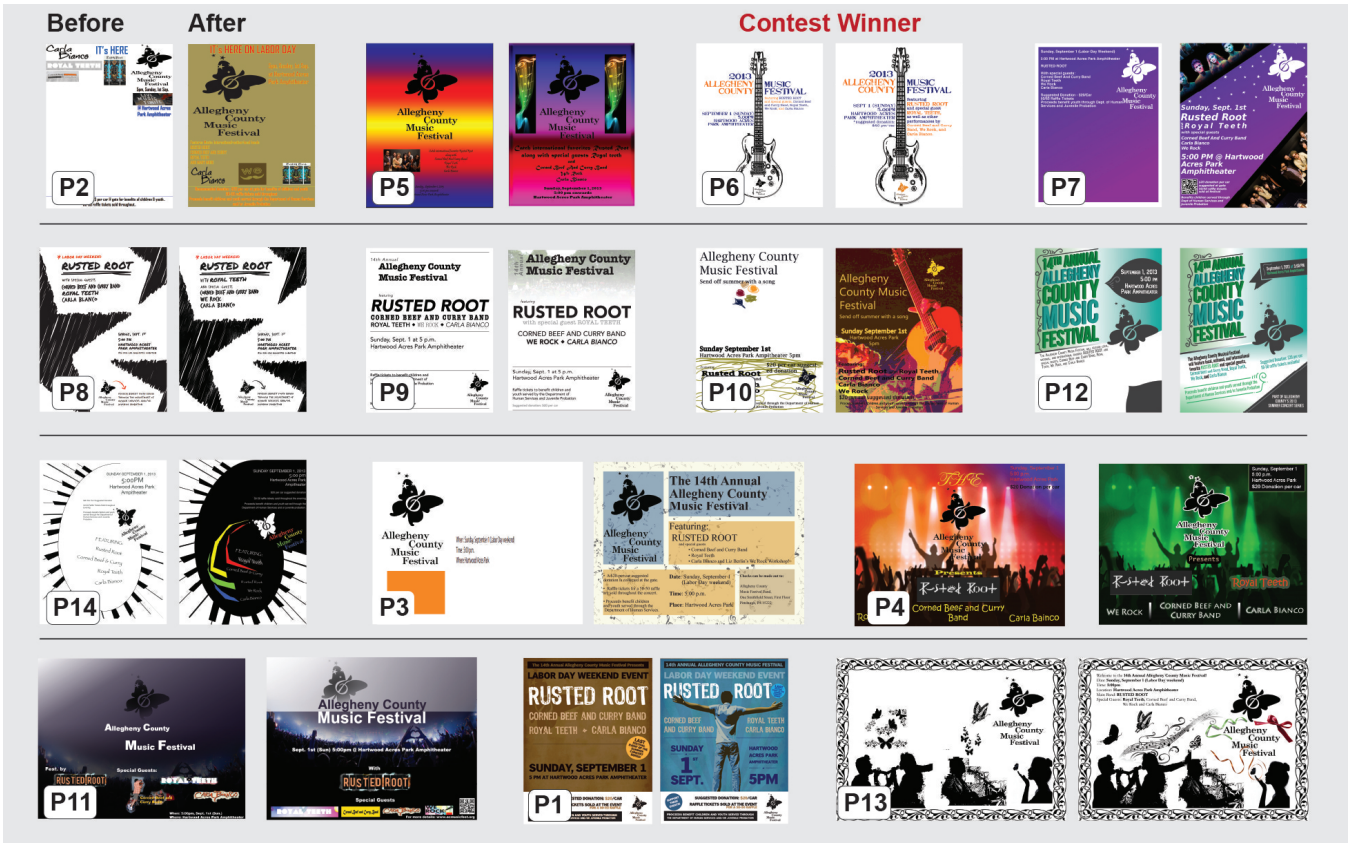


Figure 8. All drafts (left) and final designs (right) from the Study 2. P6 won the contest.

tended to disregard novice feedback because the experienced designers felt more knowledgeable. P6 said, "Not to sound pompous or anything, but I feel like I do know a little bit more than general people who don't really know much about design." Other participants felt that novice critiques were valuable in their own distinct way. P3 felt that novices provided "good emotional feedback" while experts offered "a higher-level technical critique." P1 saw novices as representative of the target audience of her design: "[Novices] are the customer, too—they are the ones that are going to the show."

Crowd vs. client critiques: Participants found both client and crowd feedback to be valuable, albeit in different ways. The crowd, as mentioned above, provided both designerly critique and audience reactions. The clients, on the other hand, were sources of domain-specific knowledge, providing input on the music festival and the goals for the poster. For our design contest, participants needed to consider both types of feedback to succeed. P5 noted that clients provided feedback specific to the festival (e.g. add a photo of the band Royal Teeth), while the crowd provided "more generic" feedback about effective design (e.g. "the font size should change"). The clients also saw their role as complementary to the crowd's. Their "general feedback" supplemented the crowd's "high-end visual feedback" (Client).

Some participants viewed the clients' lack of design expertise as problematic, and crowd critique as an exciting possible solution. They imagined "hav[ing] a crowd behind you, backing you up" (P1), i.e., using the crowd's feedback as empirical evidence to persuade clients to abandon a poor idea (P1) or embrace one favored by the designer (P3).

Crowd vs. other critique sources: Participants also compared the crowd critiques they received to other feedback sources. Participants who worked as freelancers (P1) or as the lone designer within an organization, such as a university department (P6, P8), struggled to find reliable sources of meaningful feedback beyond their clients, and these clients frequently lacked design backgrounds. Some isolated participants turned to online feedback sources, like Dribbble [18] and design forums, but were generally unsatisfied with the results. P6 praised the crowd critique for being "more specific, like it addressed very, very specific issues better than just vague 'I like it' or 'I don't like it'" comments typical of design websites. She added that the "guiding points" and "categories" of the crowd critique system seemed effective for preventing the misunderstandings and mean-spirited remarks which deterred her from these other sites.

Privacy and identity: Participants generally expressed little concern over sharing their works-in-progress with the crowd for the purposes of critique. Several mentioned that any apprehensions they might have were eased because the system was double-blind; designers were anonymous to critiquers and vice-versa. P1 said, "I like anonymous... They don't

see my name, I don't see theirs. That's good." She added that she was grateful that the system did not support direct interaction between designers and crowds, fearing that direct communication could lead to arguments and defensiveness.

How do participants use crowd critiques?
Most participants used the feedback to make changes to their design. For example, P1 increased the legibility of the top of the design and removed a distracting image; P5 changed the placement of a photo and made the background less "flashy"; and P9 changed an inappropriate typeface and added color to create visual interest.

Participants also used the feedback to decide what *not* to change. For example, P6 saw that the clients "liked how clean and simple it was so instead of changing my design to make it more flowery or whatever, I just decided to mostly keep it the way it was." Participants also mentioned situations where they agreed with feedback, but chose not to act on it. For example, P13 agreed with a crowd worker's suggestion that creating an illustration would help tie together two images, but he felt he lacked the drawing skill to execute it. The clients gave P8 negative feedback on the "stars" motif he had chosen, but he kept it because he felt he was too far along with the original concept.

Summary

This study offered evidence that crowd feedback provides value to designers across a range of backgrounds and skill levels. It demonstrated how a crowdsourcing platform could be leveraged to provide design critique for a contest scenario. Participants were generally enthusiastic about the quality, quantity, cost, and speed of crowd critiques, and found the aggregation tool straightforward and helpful. Participants also described their reactions to this novel source of critique. They compared crowds favorably to more familiar feedback sources and identified unique benefits of crowds, such as providing a broader range of perspectives, preserving limited social capital, and protecting privacy and anonymity. Finally, the study revealed how participants used crowd critiques in their design processes, weighed expertise against their own intuitions, and even used crowds to justify design decisions.

STUDY 3: IMPACT OF CROWD CRITIQUE ON DESIGN PROCESS AND RESULTS

The second study provides qualitative evidence for the efficacy of gathering crowd critique for a realistic design scenario. However, it did not directly compare crowd critique to the feedback typically available on a design contest platform. Therefore, we conducted a controlled between-subjects experiment to isolate these effects. Specifically, we investigated the following research questions:

- How do designers perceive crowd critique in comparison to the generic feedback typical on contest platforms?
- How does crowd critique affect designers' final products in comparison to generic feedback?

Method

Site
We ran our experiment on 99designs [1], a popular host of online design contests. The site has hosted nearly 300,000 contests since it was founded in 2008.

Participants
Eighteen designers (five female, ages 19–55) participated in the experiment. All but three reported their occupations as designers. They had won, on average, 5.6 contests on 99designs (min: 0; max: 33). All 18 designers were paid $10 for participating, and the winner (D18) received an additional $599 in prize money.

Experimental conditions
We randomly assigned half (nine) of the designers to the experimental condition, where they received crowd critiques via URL. The remaining nine designers, assigned to the control condition, received the following "generic" feedback: *"We checked out your design and wanted to tell you that you're on the right track. Keep up the good work!"* This type of feedback is typical of contest sites like 99designs, where many hosts provide little or no feedback to designers.

Procedure
We initiated a 99designs contest to design a poster advertising an upcoming lecture series hosted by a U.S. university. We worked with the lecture series coordinator to develop a design brief, which contained information about the lecture series and target audience, as well as more specific requirements like size and sponsor logos.

The contest lasted one week from start to finish. Participants (hereafter "designers" for sake of clarity) completed a demographics pre-survey and submitted a draft of their poster design by the end of Day 3 of the contest. On Day 4, we generated crowd critiques for all 18 designs from a total of 30 crowd workers on MTurk, and distributed these crowd critiques to designers in the experimental condition. Designers in the control condition were sent generic feedback. All designers were asked to consider their feedback, iterate, and submit a final poster design on Day 7. Pre- and post-critique versions of all 18 posters, totalling 36 unique designs, are shown in Figure 9.

Following the contest deadline, all designers completed a post-survey about their impressions of the contest and feedback. Designers who had received only the generic feedback were given the URL for the crowd critiques, which we generated on Day 4 but did not reveal until this time (Day 8). These designers completed a second post-survey about their reactions to the crowd critiques.

The final posters were evaluated by both the client (the lecture series coordinator) and crowd workers from MTurk. We sought evaluations from people with design knowledge, so we required crowd workers to score 80% or higher on a true/false design knowledge quiz that was previously validated [9]. Thirty-seven crowd workers passed the quiz and participated in the evaluation.

Figure 9. All drafts (left) and final designs (right) from Study 3. D18 won the contest.

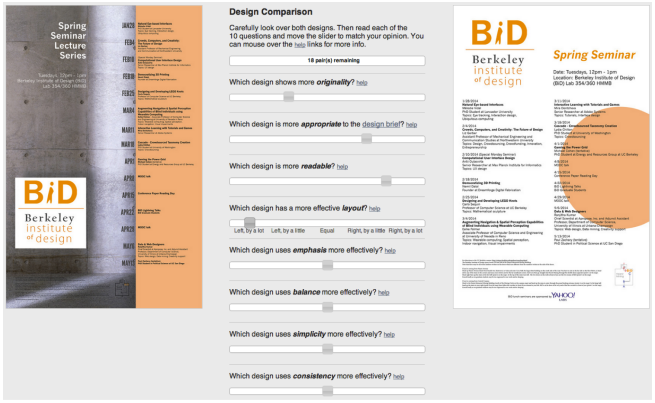


Figure 10. The rating interface used by crowd workers and the client to evaluate the 18 poster designs along 10 dimensions.

Raters viewed both poster versions from all 18 participants (36 designs total), unlabeled and in random order. They compared two versions across 10 dimensions: adherence to the seven design principles, originality, similarity, and overall quality (see Figure 10). Raters moved a slider left or right, more towards one design or the other, to indicate which design embodied each dimension more effectively.

Analysis

To analyze the non-parametric survey data, we ran independent samples Mann-Whitney U-tests. We also analyzed designers' responses to open-ended questions about the feedback and share illustrative quotes below.

To analyze the posters, we first calculated intraclass correlation (ICC) scores to establish the reliability of the crowd's ratings (see Table 1. Eight of the 10 dimensions scored 0.75 or higher (considered "excellent" on a scale of 0–1), and the remaining two dimensions, Balance and Appropriateness, scored 0.60 or higher (considered "good"). Thus, there was high agreement among crowd raters.

For each designer, we computed the average change in score between his or her initial and final designs (delta), as rated by the crowd workers on a 0–10 Likert scale. The maximum delta was ±5 points, except for Similarity, which was a 0–10 scale. Positive deltas indicate the post-feedback design improved along that dimension.

We ran independent samples t-tests comparing the crowd feedback deltas across all 10 dimensions to the generic feedback deltas. these deltas across conditions. We also examine poster ratings by design experience. For the latter analysis, we divided the participant pool into thirds according to the number of 99designs contests they had won.

Figure 11. Post-survey results. Only the first three questions (marked with *) have statistically significant different responses between conditions ($p<0.05$).

Results and discussion

Survey responses are shown in Figure 11, and poster evaluation results are shown in Table 1. In general, designers rated the feedback from CrowdCrit higher than the generic feedback. Designs improved along most dimensions between the initial and final iterations, but there was no significant difference between conditions.

Designers receiving crowd crits felt they noticed more issues

Designers who got crowd critiques reported noticing significantly more issues they would have ordinarily missed (6.13 vs. 3.67, $p<0.05$). Some attributed this to the higher quantity and variety of critiques available on CrowdCrit. D10 wrote, "I didn't realise that in a design there are so many points to look for... now I know what to look for in my future designs." D2 wrote, "The feedback about lacking contrast, imagery and the bottom text being too small were right on. (I can't believe I didn't see it myself)."

Designers receiving crowd crits felt they made made better designs, but third-party ratings show little difference

Compared to generic feedback, designers felt that crowd critiques helped them make a significantly better overall design (6.25 vs. 3.67, $p<0.05$). Many provided specific examples of crowd feedback that led them to make improvements. D11 described how, after reviewing critiques in the Balance tab, he "noticed that the bottom part of my entry lacks something to balance the whole thing... so I added edited photos of the university to fill in the gap." D15 described "changing the layout of the dates to make it less confusing", and D13 "[c]hanged a lot of things that weren't making sense at all like using dark colour fonts under the dark background."

However, despite designers' greater sense of improvement in the crowd feedback condition, the third-party ratings show no significant differences between conditions for nine of the 10 dimensions. Only final designs receiving generic feedback scored significantly higher on Simplicity than designs getting crowd feedback (+0.63 vs. +0.09, $p<0.05$). Designs tended to improve modestly ($\leq$1 point) across all dimensions following either type of feedback.

Dimension	ICC score	Crowd avg. Δ	Generic avg. Δ
*Simplicity	0.94‡	+0.09	+0.63
Balance	0.71†	+0.44	+0.67
Consistency	0.84‡	+0.25	+0.40
Appropriateness	0.64†	+0.20	+0.26
Readability	0.88‡	+0.74	+0.81
Layout	0.86‡	+0.65	+0.69
Emphasis	0.88‡	+0.54	+0.52
Originality	0.95‡	+0.20	+0.12
Similarity	0.99‡	6.51	6.19
Quality	0.84‡	+0.69	+1.00

Table 1. Average change (Δ) between pre-critique and post-critique designs for the 10 dimensions evaluated. Only Simplicity (marked with *) was significantly different between conditions ($p<0.05$). ICC scores with ‡ have "excellent" agreement; scores with † have "good" agreement.

Designers receiving crowd crits felt they made bigger changes, but ratings show only inexperienced designers did

The third significant survey result helps to explain the discrepancy between designers' perceptions and the poster ratings. Compared to generic feedback, designers reported making significantly more changes to their designs as a result of crowd critiques (5.25 vs. 2.89, $p<0.05$). Some mentioned that the higher specificity and concreteness of crowd critiques made it easier to recognize problems and make changes. D1 appreciated that crowd feedback "showed where the change was needed...in a graphical manner." D4 wrote, "[Crowd] feedback was precise, detailed and to the point. Even though it didn't favor my design it was better than someone saying they like it, or nice and not giving any more details."

The poster ratings partially support these claims. As mentioned above, there was no significant difference in Similarity when designers of all experience levels are analyzed together. Separating the analysis by designers' contest experience, however, tells a different story. High-experience designers who received crowd critique produced more similar final designs (8.47 vs. 8.01, $p<0.05$), while low-experience designers getting crowd critique produced less similar final designs (4.81 vs. 5.44, $p<0.05$). As these means suggest, inexperienced designers made bigger changes between iterations, regardless of feedback type. Generic feedback was correlated with higher Simplicity scores, regardless of experience ($p<0.05$). There were no other significant differences across dimensions.

IMPLICATIONS

Evidence from the first study suggested that aggregated crowd critiques can address many of the issues raised by expert critique providers. Designers from the 99designs contest site expressed enthusiasm for crowd critiques and offered concrete examples of how they influenced their design process, echoing the findings of our second study. Compared to the generic feedback condition, designers who got crowd critiques reported significantly better recognition of issues, more changes, and better overall designs. By mitigating "good subject" effects [19] through our controlled experiment, these results show that CrowdCrit provides value over the generic feedback typical in design contests.

Yet, when we examine independent ratings of their designs, we see no significant difference in quality or most other dimensions. One interpretation of these results is that crowd critique leads designers to think they are making significant improvements, when in fact, they tend to make minor revisions. The opportunity to respond to specific issues, provided by CrowdCrit, may fixate designers on addressing minor, easily-addressed problems rather than contemplating broader, more substantive changes. The aggregation interface provides a "Top Feedback" section aimed to mitigate this problem, but this orientation towards simple frequency may have oriented designers towards issues that were popular or easily identified (e.g. a typo) rather than important. Further, presenting critiques in list form may have led to changes that were idiosyncratic, rather than contributing to a more synergistic improvement. This could have resulted in designs that were "busier" but not necessarily better, possibly explaining why crowd critique led to more complex designs. Our first study, which found that many designers approached their revisions by working their way through the list of critiques, also supports this interpretation.

Another partial explanation may come from the nature of our study site, a host of online design contests. Typically only the winner is paid, so designers may find that submitting multiple revisions to fewer contests is a less effective strategy than submitting a single reasonable version to many contests. The culture of 99designs may have discouraged designers from making significant changes, even if the crowd critiques were considered valuable. Our finding that designers with more contest experience made fewer changes, regardless of critique type, also supports this interpretation. Crowd critique led inexperienced designers to make bigger changes, possibly because they had not yet embraced this "shotgun approach" as more productive.

Systems like CrowdCrit may be most appropriate for domains in which quality is somewhat subjective, experts frequently disagree, and a rich diversity of feedback is valued. Although usability heuristics have now been widely adopted as a foundation of HCI practice, Nielsen and Molich found little correlation between evaluators, experts adjusted their rubrics in response to novice feedback, and the average novice found only 20–50% of issues identified by experts. They acknowledge that usability evaluation is challenging, yet argue convincingly that it is still worthwhile because "even finding some problems is of course better than finding no problems" [20]. We propose that a similar argument may apply to crowdsourced design critique and possibly other domains of crowdsourced feedback.

FUTURE WORK

Enhance scaffolding

Scaffolding allows novices to quickly become productive within an unfamiliar domain to gradually learn to do new things without support. As such, scaffolding can come in many forms. Our critique statements provided a first approximation at scaffolding visual design critique, but we did not measure the effect on crowd worker learning, or implement techniques such as progressively "fading" scaffolds, requiring the learner to draw on memory to fill new gaps [25]. Crowd feedback systems could be modified to monitor critique providers and, as they gain experience, remove supports for structured feedback. For example, CrowdCrit could adapt the pre-authored critique statements to match a worker's skill level, perhaps starting with surface-level elements (like fonts, text, and images), moving to more abstract principles (like balance and emphasis), eventually dropping all of these in favor of more open-ended critique.

Support new principles

For crowd feedback services to generalize across domains, they should offer flexibility in *how* they structure novice behavior. For CrowdCrit, this could mean allowing designers to author their own set of visual design principles or to modify the ones provided in our system. This feature could also make CrowdCrit useful in other domains, such as interface design or architectural design, and to other types of crowds, such as user communities or concerned citizens.

Explore new power dynamics

This paper focused on the practical scenario of introducing feedback in design contests. As a result, we had the opportunity to observe interactions between designers, crowds, and clients (or contest hosts). The introduction of crowd critique created the potential to challenge the traditional relationship between the designer and client. Instead of merely responding to the clients' whims, designers saw an opportunity to use crowd feedback as empirical support for certain design decisions. The crowd, in a sense, represents both the target audience and a body of knowledgeable designers. Future work on crowd feedback systems can explore new interaction models where clients also view and interpret crowd feedback.

CONCLUSION

We presented CrowdCrit, a system which uses crowdsourcing to provide designers with a fast, scalable, and reasonably priced source of feedback. We built a working prototype of CrowdCrit and deployed it in three studies. This paper makes the following contributions. First, we present the CrowdCrit system itself, demonstrating how to structure the complex task of design critique for crowd workers, and how to aggregate and present large, diverse quantities of crowd feedback for review. Second, we compared crowd and expert critiques, showing how aggregated crowd critiques converges towards expert feedback, and demonstrating that agreement trends low among both novice and expert critiquers. Our second study used interviews to provide rich descriptions of a diverse group of designers' attitudes, concerns, and enthusiasm towards crowd feedback and patterns of use in a real-world design context. Our third study showed that crowd feedback led designers to report noticing more issues, enacting more changes, and producing better designs that the generic feedback more typical of online contest hosts. It also identified a gap between designers' perceptions and independent ratings of changes to their final designs. These evaluations are among the first to attempt to describe and quantify the value of crowd feedback. Finally, we discuss design implications

for building crowd feedback systems that provide valuable and equitable services to both designers and crowd workers.

ACKNOWLEDGMENTS

The authors wish to thank our designer participants, crowdworkers, Allegheny County (PA) Parks Department, 99designs.com staff, and our anonymous reviewers. Financial support provided by the National Science Foundation under IIS grants 1210836, 1208382, and 1217096.

REFERENCES

1. 99designs. `http://www.99designs.com/`.
2. Barrett, T. A comparison of the goals of studio professors conducting critiques and art education goals for teaching criticism. *Studies in art education* (1988), 22–27.
3. Bransford, J. D., Brown, A. L., and Cocking, R. R., Eds. *How people learn: brain, mind, experience, and school*. National Academy Press, Washington, D.C., 2000.
4. Dannels, D., Gaffney, A., and Martin, K. Beyond content, deeper than delivery: What critique feedback reveals about communication expectations in design education. *Int. J. Schol. Teach. & Learn. 2*, 2 (2008).
5. Dannels, D. P., and Martin, K. N. Critiquing critiques a genre analysis of feedback across novice to expert design studios. *Jo. Bus. & Tech. Comm. 22*, 2 (2008), 135–159.
6. Dondis, D. A. *A primer of visual literacy*. MIT Press, Cambridge, Mass., 1973.
7. Dow, S., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D., and Klemmer, S. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proc. CHI 2011* (2011), 2807–2816.
8. Dow, S., Gerber, E., and Wong, A. A pilot study of using crowds in the classroom. In *Proc. CHI 2013*, CHI '13 (New York, NY, USA, 2013), 227–236.
9. Dow, S. P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L., and Klemmer, S. R. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact. 17*, 4 (Dec. 2010), 18:1–18:24.
10. Easterday, M. W., Rees Lewis, D., Fitzpatrick, C., and Gerber, E. M. Computer supported novice group critique. In *Proc. of DIS '14* (2014).
11. Feedback army. `http://www.feedbackarmy.com/`.
12. Feldman, E. *Practical Art criticism*. Prentice Hall, Englewood Cliffs, NJ, 1994.
13. Five second test. `http://www.fivesecondtest.com/`.
14. Forrst. `http://www.forrst.com/`.
15. Klemmer, S. R., Hartmann, B., and Takayama, L. How bodies matter: five themes for interaction design. In *Proceedings of DIS*, ACM (2006), 140–149.
16. Kulkarni, C., and Klemmer, S. R. Learning design wisdom by augmenting physical studio critique with online self-assessment. Tech. rep., Stanford U., 2012.
17. Lidwell, W., Holden, K., and Butler, J. *Universal principles of design*. Rockport Pub, 2010.
18. Marlow, J., and Dabbish, L. From rookie to all-star: Professional development in a graphic design community of practice. In *Proc. CSCW 2014* (2014).
19. Nichols, A. L., and Maner, J. K. The good-subject effect: investigating participant demand characteristics. *The Journal of general psychology 135*, 2 (Apr. 2008), 151–165. PMID: 18507315.
20. Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, ACM (New York, NY, USA, 1990), 249–256.
21. Reynolds, G. *Presentation Zen: Simple ideas on presentation design and delivery*. New Riders, 2011.
22. Sadler, D. R. Formative assessment and the design of instructional systems. *Instr. Sci. 18*, 2 (1989), 119–144.
23. Schön, D. *Educating the Reflective Practitioner*. Jossey–Bass Publishers, San Francisco, 1990.
24. Shneiderman, B. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. VL/HCC 1996* (1996), 336–343.
25. Soloway, E., Guzdial, M., and Hay, K. E. Learner-centered design: the challenge for HCI in the 21st century. *interactions 1* (April 1994), 36–48.
26. Tinapple, D., Olson, L., and Sadauskas, J. Critviz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology 15*, 1 (2013), 29.
27. Tohidi, M., Buxton, W., Baecker, R., and Sellen, A. Getting the right design and the design right. In *Proceedings of CHI*, ACM (2006), 1243–1252.
28. Wattenberg, M., and Kriss, J. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 549–557.
29. Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 227–236.
30. Willett, W., Heer, J., Hellerstein, J., and Agrawala, M. Commentspace: structured support for collaborative visual analysis. In *Proc. CHI 2011* (2011), 3131–3140.
31. Williams, R. *The Non-Designer's design book*. Peachpit Press, 2008.
32. Xu, A., and Bailey, B. What do you think?: a case study of benefit, expectation, and interaction in a large online critique community. In *Proc. CSCW 2012* (2012), 295–304.
33. Xu, A., Bailey, B. P., and Huang, S.-W. Voyant: Generating structured feedback on visual designs using a crowd of non-experts. In *Proc. CSCW 2014* (2014).