A Classroom Study of Using Crowd Feedback in the Iterative Design Process

Anbang Xu[1], Huaming Rao[2], Steven P. Dow[3], Brian P. Bailey[4]
[1] IBM Research - Almaden, San Jose, CA, USA, anbangxu@us.ibm.com
[2] Nanjing University of Science and Technology, Nanjing, China, huaming.rao@gmail.com
[3] Carnegie Mellon University, Pittsburgh, PA, USA, spdow@cs.cmu.edu
[4] University of Illinois, Urbana, IL, USA, bpbailey@illinois.edu

ABSTRACT

Crowd feedback systems offer designers an emerging approach for improving their designs, but there is little empirical evidence of the benefit of these systems. This paper reports the results of a study of using a crowd feedback system to iterate on visual designs. Users in an introductory visual design course created initial designs satisfying a design brief and received crowd feedback on the designs. Users revised the designs and the system was used to generate feedback again. This format enabled us to detect the changes between the initial and revised designs and how the feedback related to those changes. Further, we analyzed the value of crowd feedback by comparing it with expert evaluation and feedback generated via free-form prompts. Results showed that the crowd feedback system prompted deep and cosmetic changes and led to improved designs, the crowd recognized the design improvements, and structured workflows generated more interpretative, diverse and critical feedback than free-form prompts.

Author Keywords

Crowd feedback; crowdsourcing; design iteration; visual design; creativity.

ACM Classification Keywords

H.5.3 [Information Interface and Presentation]: Group and Organization Interfaces – Collaborative computing.

INTRODUCTION

Feedback is essential for the iterative design process because it reveals gaps between what is intended by the designer and how an audience interprets the design [15]. Knowing such gaps can help the designer iterate toward an outcome that better connects with its target audience.

However, receiving quality feedback can be hard. Face-to-face critiques are the gold standard [13] but impose an organizational burden. Peer requests are quick and simple but burn social capital and the feedback may be biased by friendship or competition [30]. Online communities provide another outlet but the quantity and quality of feedback typically falls below users' expectations [33, 35].

To overcome these issues, researchers have been exploring how *crowd feedback systems* can leverage online crowds as a simulated audience to provide feedback on designs. Researchers have used online crowds to collect preferences on design alternatives [2, 11] and to generate structured feedback on individual designs [23, 36]. Crowd-based usability sites [39, 40, 42] can be also utilized to conduct surveys and task-oriented tests of Web designs.

Although this class of system shows promise, there is little empirical knowledge of the effectiveness of these systems for iterative design. Prior work has shown that crowd feedback helps users discover problems with their designs [36], but this finding was based on user perceptions of the feedback and did not study changes to the designs. Other studies have shown the validity of using online crowds to test interfaces based on task performance [21] and for conducting A-B tests of design alternatives [11]. However, none of these studies investigated how crowd feedback prompts changes to a design, the validity of the feedback relative to expert evaluation, or how structuring the feedback generation process affects feedback content.

To fill this empirical gap, we conducted a study of how crowd feedback affects visual design. Crowd feedback was generated with a research prototype called Voyant [36]. Voyant generates five types of feedback on a visual design (e.g., first impressions, adherence to design guidelines, etc.) by decomposing the generation process into micro-tasks that can be executed by crowd workers without design expertise. The study was conducted in context of a two-week project in an introductory visual design course. Users (N = 10) created initial designs satisfying a challenging design brief prepared by the instructor. Crowd feedback was generated on the designs and given to the users. The users then revised the designs for their final deliverable. After the deadline, crowd feedback was once again generated so that we could examine changes between the initial and revised designs and how the feedback affected those changes. We also compared the feedback content to expert evaluation and feedback generated via open-ended

CSCW '15, March 14 - 18 2015, Vancouver, BC, Canada
Copyright 2014 ACM 978-1-4503-2922-4/15/03…$15.00
http://dx.doi.org/10.1145/2675133.2675140

prompts (e.g., what do you think of this design?). The main findings are:

- Crowd feedback was able to prompt a variety of changes in the iterative design process. A majority of the changes reported for the designs were deep changes relating to the theme and layout of a design. Cosmetic changes were also reported relating to the colors and fonts used in a design.
- The non-expert crowd was able to reliably evaluate design improvements based on design guidelines, as compared against expert ratings. However, the non-expert crowd and domain experts did not agree on how well a design achieved its communicative goals.
- The discourse in the structured feedback generated with the system, which poses specific prompts to the crowd, was more diverse and detailed than the discourse in the free-form feedback generated without such prompts. The most common category of critique discourse in the structured feedback was *interpretation* whereas the most common category in free-form feedback was *judgment*. These differences, along with the results from linguistic style analysis, show that structured workflows can help crowd workers think more critically about a design.

These outcomes provide initial evidence that can increase the design community's confidence in and knowledge of using crowd feedback systems in the iterative design process. The outcomes also provide insights into workflow patterns for improving the generation of crowd feedback.

RELATED WORK

We describe how our work contributes to the class of crowd feedback systems and situate our work within prior studies of design feedback and crowdsourcing in design.

Crowd Feedback Systems

The purpose of crowd feedback systems is to enable users to receive helpful, timely, and affordable feedback on their designs. For example, FeedbackArmy [39] is a commercial system that enables a user to pose open-ended questions about a design and the site returns free-form responses collected from an online crowd. Similar sites include Usabilla [42], Fivesecondtest [40], and UITests [41]. In research, Voyant is a crowd feedback system that generates five types of feedback by posing specific prompts to the crowd [36]. The rationale for the feedback types was based on a need finding study conducted with designers at various skill levels. CrowdCrit is similar, but uses learning theory to scaffold worker responses for feedback generation [23].

As these technologies emerge, it is important to assess their benefits and limits for design. Prior studies have shown that crowd feedback helps users discover problems with their designs [36] and is an enjoyable part of the process [11]. However, prior work has not investigated how crowd feedback prompts changes in the iterative design process. Our work addresses this gap. From a classroom study, we report the types and depth of changes prompted by crowd feedback for visual design and how this feedback compares to expert evaluation. We also compare the content of the structured crowd feedback, which uses specific prompts, to free-form feedback collected from the same online crowd.

Prior crowdsourcing studies showed that directing worker attention with prompts and using examples improves the responses from the crowd for data analysis [20, 33]. Our works extends a similar concept for visual design and compares it to crowd feedback elicited holistically.

Studies of Design Feedback

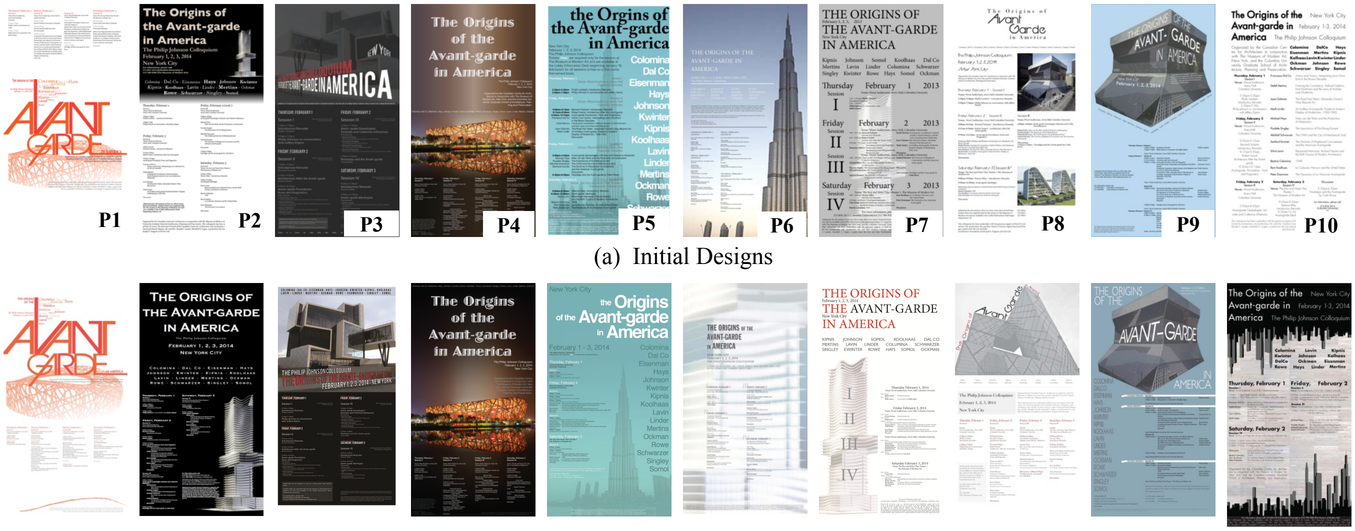
Studies show that feedback from designers' peers and instructors improves their ability to create effective designs and understand how to better assess creative work [16, 27]. Designers who receive feedback during iterative design produce higher quality outcomes than those who do not [12, 13]. Peer feedback can also foster effective communication and collaboration among designers [9, 22]. Moreover, researchers have developed a typology of critique discourse generated by design novices and experts [8]. Feedback from experts was more collaborative and interpretative and less directive than feedback from novices.

Another body of work has investigated the effects of audience feedback in the design process. The feedback helped designers iterate toward design solutions that better connect with the intended audience [15, 32]. For example, to better design a computer-based system, designers often present their designs to a target audience and perform qualitative studies such as walkthroughs with a small group of individuals to collect feedback on the design [5].

Crowd feedback systems purportedly allow designers to receive similar feedback, acquired from a larger and more diverse audience, faster, and with lower cost. But how does a designer react to the feedback delivered from an online crowd, how does an online crowd react to the changes made by the designer, and how does crowd feedback compare to expert evaluation and typologies of critique discourse? Our work studies these questions for one crowd feedback system in context of visual design.

Crowdsourcing in Design

Researchers have investigated many directions for how a non-expert crowd can aid design activities. For example, Yu et al. showed how crowds can be integrated with genetic algorithms [38] and analogical transfer [37] to generate creative ideas. For example, a crowd can generate better ideas when given appropriate analogical schemas. In the CvC design method, the crowd works with the designer to form a team as part of an open design competition [25]. The designer can leverage these team members to brainstorm design solutions. In our work, we are leveraging online crowds for the purpose of providing feedback on visual designs and studying how it affects the design process.



(a) Initial Designs

(b) Revised Designs

Figure 1. The designs created during the study. (a) Initial Designs: preliminary designs completed by participants in the first week. (b) Revised Designs: after receiving crowd feedback, participants spent one week revising their initial designs. Each column represents the initial / revised pair of designs for a participant.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Theme			●			●		●		
Layout	●	●			●	●	●	●		●
Typeface		●	●		●			●		●
Color		●	●	●	●	●	●			

Table 1. The distribution of the four types of changes made by participants according to crowd feedback.

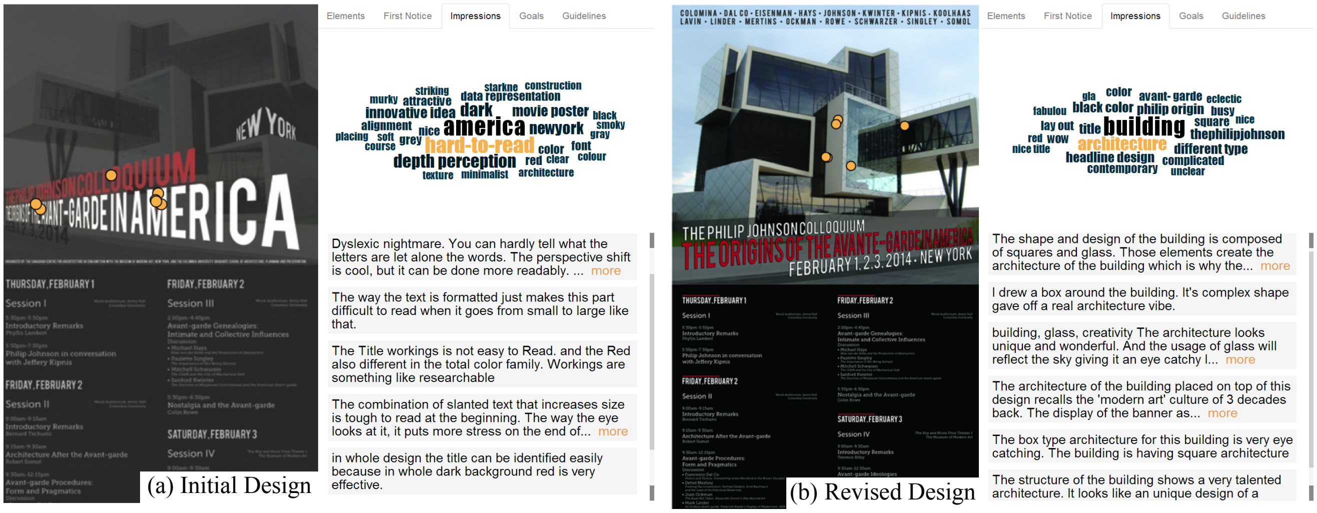
VOYANT: THE CROWD FEEDBACK SYSTEM USED

Voyant is a system that engages an online crowd as a simulated audience to collect, aggregate, and present their interpretation of a design [34, 36]. To use the system, a user uploads a design, configures the audience (age, gender, geography), and submits the design for feedback generation.

The system generates five types of feedback: *Elements* are the individual elements visible or otherwise perceived in the design including colors, shapes, objects, and activities. *First Notice* is the self-reported order in which elements are seen in the design. *Impressions* are the first perceptions formed in one's mind upon viewing the design. *Guidelines* refer to how well the design is perceived to meet guidelines in the domain using a seven point scale. The guidelines currently include proximity, alignment, repetition and contrast, which are commonly taught in visual design. *Goals* is how well the design is perceived to meet its communicative goals on a seven point scale. If this type of feedback is selected, the user is prompted to briefly describe each of her goals.

The feedback generation process is decomposed into a set of micro-tasks doable by crowd workers without design expertise. The micro-tasks relate to a description and interpretation phase. The first phase enumerates what elements can be "seen" in a design. The second phase is to interpret the design related to the visual hierarchy, first impressions, guidelines, and communicative goals entered by the user. In each phase, a micro-task focuses worker attention on a specific aspect of a design related to the feedback type rather than soliciting holistic judgments. For each task, a worker reacts to a prompt (e.g. what is your first impression?), annotates the design to indicate the regions associated with her reaction, and enters brief rationale. Using the coordinates of the annotation, the system also provides up to three elements (e.g. colors, shapes, or objects) identified by prior workers as examples to aid the rationale. The responses from each worker are then aggregated. The micro-tasks are submitted to Amazon Mechanical Turk, a popular online labor market. To reduce latency and costs, the system does not currently use explicit quality controls such as peer-assessment. It also allows a worker to perform multiple micro-tasks. The full set of feedback requires about $10 and several hours to generate.

Once complete, each feedback type is presented as a visual summary of the crowd's reactions and annotations. The system also provides interactions for exploring details of the feedback from the perspective of either the reactions or annotations on the design. We chose Voyant for our study



Dyslexic nightmare. You can hardly tell what the letters are let alone the words. The perspective shift is cool, but it can be done more readably. ... more

The way the text is formatted just makes this part difficult to read when it goes from small to large like that.

The Title workings is not easy to Read. and the Red also different in the total color family. Workings are something like researchable

The combination of slanted text that increases size is tough to read at the beginning. The way the eye looks at it, it puts more stress on the end of... more

in whole design the title can be identified easily because in whole dark background red is very effective.

The shape and design of the building is composed of squares and glass. Those elements create the architecture of the building which is why the... more

I drew a box around the building. It's complex shape gave off a real architecture vibe.

building, glass, creativity The architecture looks unique and wonderful. And the usage of glass will reflect the sky giving it an eye catchy l... more

The architecture of the building placed on top of this design recalls the 'modern art' culture of 3 decades back. The display of the banner as... more

The box type architecture for this building is very eye catching. The building is having square architecture

The structure of the building shows a very talented architecture. It looks like an unique design of a

Figure 2. The *Impression* feedback on the initial and revised designs for the participant [P3]. (a) The impression word "hard to read" on the initial design prompted the participant to change the image treatment. The revision (b) was no longer perceived as "hard to read" by the crowd.

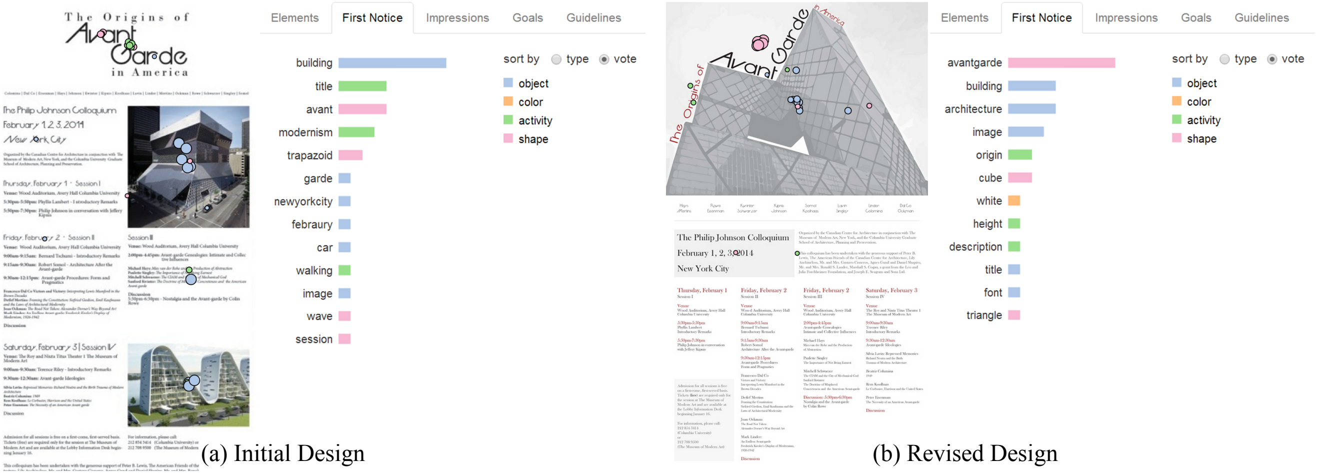


Figure 3. In (a), the first noticed element by the crowd is "building." In the revision (b), the first noticed element is the event. The feedback helped the participant [P8] understand how to utilize an image to attract people's attention more effectively.

because it is representative of the class of crowd feedback systems for visual design.

RESEARCH QUESTIONS

The purpose of the study was to understand how the use of crowd feedback affects the iterative design process. As a first step, the study addressed three related research questions:

RQ1: How is crowd feedback leveraged by users to iterate on a visual design? What types of changes are prompted by the feedback? What is the depth of the changes?

RQ2: Once users iterate on a design, how effectively does the crowd recognize the changes made to the design? What is the validity of the crowd feedback, e.g. how do the crowd's ratings compare to the ratings of experts for guidelines and goals?

RQ3: How is crowd feedback affected by structuring the generation process with prompts? What are the differences and overlaps between structured and free-form feedback?

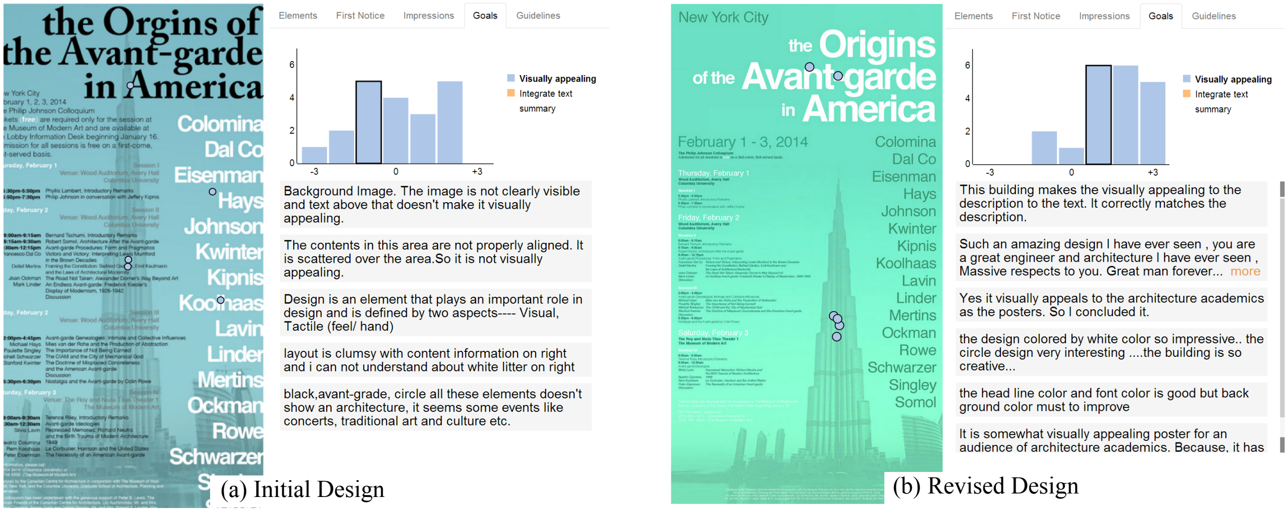
These questions are not exhaustive, but do provide a starting point for understanding the effects of crowd feedback in the context of iterative design and identifying opportunities for improving this class of system.

METHOD

To answer the research questions, we deployed the crowd feedback system in an entry-level visual design course at a large private university in the United States.

Participants

Ten students volunteered to participate in the study out of the fifteen students enrolled in the course. Students came from various disciplines including engineering, computer science, and information technology and they were trying to gain experience with visual design. The third author was a co-instructor of the course and only invited participation during the study to minimize conflicts of interest.



(a) Initial Design

Background Image. The image is not clearly visible and text above that doesn't make it visually appealing.

The contents in this area are not properly aligned. It is scattered over the area.So it is not visually appealing.

Design is an element that plays an important role in design and is defined by two aspects---- Visual, Tactile (feel/ hand)

layout is clumsy with content information on right and i can not understand about white litter on right

black,avant-grade, circle all these elements doesn't show an architecture, it seems some events like concerts, traditional art and culture etc.

(b) Revised Design

This building makes the visually appealing to the description to the text. It correctly matches the description.

Such an amazing design I have ever seen , you are a great engineer and architecture I have ever seen , Massive respects to you. Great man forever... more

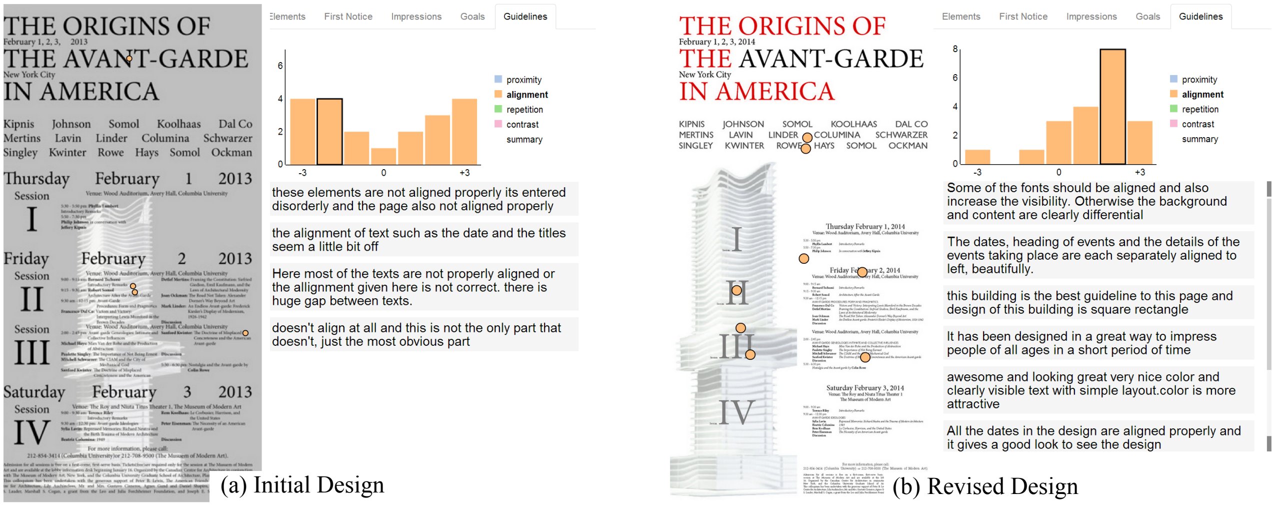
Yes it visually appeals to the architecture academics as the posters. So I concluded it.

the design colored by white color so impressive.. the circle design very interesting ....the building is so creative...

the head line color and font color is good but back ground color must to improve

It is somewhat visually appealing poster for an audience of architecture academics. Because, it has

Figure 4. In (a), the design goal "visually appealing" is not well supported. After the participant [P5] changed the layout and typeface, the average rating on the design goal in (b) was improved from 0.05 (σ = 1.5) to 1.5 (σ = 1.2). The range of the rating scale shown is -3 (low) to 3 (high).



(a) Initial Design

these elements are not aligned properly its entered disorderly and the page also not aligned properly

the alignment of text such as the date and the titles seem a little bit off

Here most of the texts are not properly aligned or the allignment given here is not correct. there is huge gap between texts.

doesn't align at all and this is not the only part that doesn't, just the most obvious part

(b) Revised Design

Some of the fonts should be aligned and also increase the visibility. Otherwise the background and content are clearly differential

The dates, heading of events and the details of the events taking place are each separately aligned to left, beautifully.

this building is the best guideline to this page and design of this building is square rectangle

It has been designed in a great way to impress people of all ages in a short period of time

awesome and looking great very nice color and clearly visible text with simple layout.color is more attractive

All the dates in the design are aligned properly and it gives a good look to see the design

Figure 5. The *Guidelines* feedback on the participant's designs [P7]. Based on the feedback, the participant changed the layout and color in (a). The average rating on the guideline *alignment* in (b) was improved from -1.0 (σ = 2.3) to 1.3 (σ = 1.4).

Design Project

The course contained a two-week project where the students needed to design an event poster for an architecture conference named "Origins of the Avant-garde in America." All students enrolled in the course were assigned the project and produced an initial and revised design. However, only those students who participated in the study received crowd feedback from our system and only their designs were included in any of the analyses.

Procedure

Our study was conducted in context of the two-week project. In the first week, each participant created an initial design of a poster for the event (*Initial Design*, see Figure 1a). Participants electronically sent their designs to the research team along with a description of their communicative goals for the poster. Since participants were assigned the same project, most of the stated goals were the same – to attract the audience's attention and provide details of the event. Other goals were more specific (e.g. "*Lay out the Thursday-Saturday events in a pleasant way*", "*Integrate text and image smoothly…*"). The research team generated the complete set of crowd feedback on the designs and notified the participants that the feedback was available. Participants then had one more week to revise their designs (*Revised Design*, see Figure 1b) and submit it. Feedback was generated on the revised designs in order to compare it to the feedback on the initial designs. After submitting the Revised Designs, participants completed a survey (see Measures) for which they received a $3 gift card.

The system recruited workers from Mechanical Turk. Workers were paid five cents (US) per task. The costs were

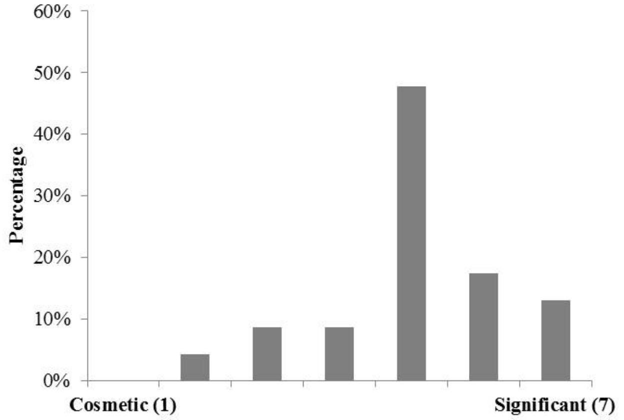


Figure 6. Distribution of ratings of the depth of changes made on the Initial Designs (n = 22).

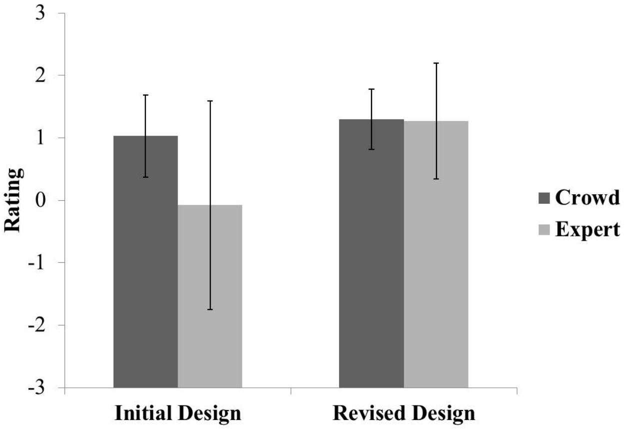


Figure 7. Overall ratings of the designs from the non-expert crowd and experts on the Initial and Revised designs.

paid by the research team. The feedback was generated by about 500 workers within 24 hours. We note that about 67% of workers performed two or more micro-tasks for feedback generation while about 14% of the workers provided feedback on both the Initial and Revised Designs.

After the two-week project ended, we collected free-form feedback from crowd workers on the same designs in order to compare it against the structured feedback. The free-form feedback was not shared with the participants. The payment for a task in the free-form condition was the same as for a task in the system. Though the workloads may differ, prior work shows the payment affects wait time, not the response quality [17, 24]. In contrast to the structured (prompted) feedback generated by the system, workers in the free-form condition responded to a general, open-ended question: "What do you think about the design and why do you think that is?"

Measures

In addition to the designs and the crowd feedback collected on those designs, the study collected four sets of measures:

Survey responses. On the survey, a participant described each notable change made to her design based on the crowd feedback and estimated the depth of the change on a seven-point scale ranging from "Cosmetic"(1) to "Significant" (7). Each participant also rated to what degree the feedback impacted their revision overall and how helpful each feedback type was on a seven-point scale. The survey responses were used to answer our first research question.

Expert evaluations. Three experts in visual design were recruited to evaluate the collected designs based on the same criterion used by crowd workers. In the crowd feedback system used, twenty crowd workers rated how well a design adheres to each of the four guidelines (proximity, alignment, repetition, and contrast) and each of the goals described by participants on a seven-point scale. Similarly, the experts rated how well the designs met each of the guidelines and goals on the same scale. The designs were presented in a random order. The *overall rating* of a design was calculated as the mean of the ratings on the guidelines and goals from the experts. Inter-rater reliability was measured by the correlation coefficients between the ratings provided by the experts. The correlations ranged from 0.61 to 0.72, with an average correlation of 0.67. Comparing the ratings between experts and non-experts would allow us to assess the validity of feedback generated by a non-expert crowd, which was our second research question.

Content analysis. For the structured and free-form feedback, we measured the feedback genres, topic diversity, and linguistic styles. For feedback genres, we adopted the categories of critique discourse developed in prior work [8] and coded a large sample of the feedback to identify what is discussed. For topic diversity, we applied a standard topic modeling method – the latent Dirichlet allocation (LDA) to extract the topics [4]. For linguistic styles, the Linguistic Inquiry Word Count (LIWC) program was used to extract psycholinguistic features from the content. Results of the analysis were used to answer our third research question.

TYPES OF CHANGES (RQ1)

Following qualitative analysis methods [29], the changes on the Initial Designs described in the survey were coded using a bottom-up approach. The descriptions of the changes were first segmented into the smallest logical units. A first pass was then performed to assign categories to the units and subsequent passes were made to revise and aggregate the categories. We found that there were four types of design changes (see Table 1):

- **Theme**. Change the visual theme or main image in a design.
- **Layout**. Reorganize visual elements in a design.
- **Typeface**. Change font style and size in a design.
- **Color**. Change the color scheme in a design.

We found that significant changes were often related to the theme and layout of a design, while cosmetic changes were often related to the color and font in a design. Note that the

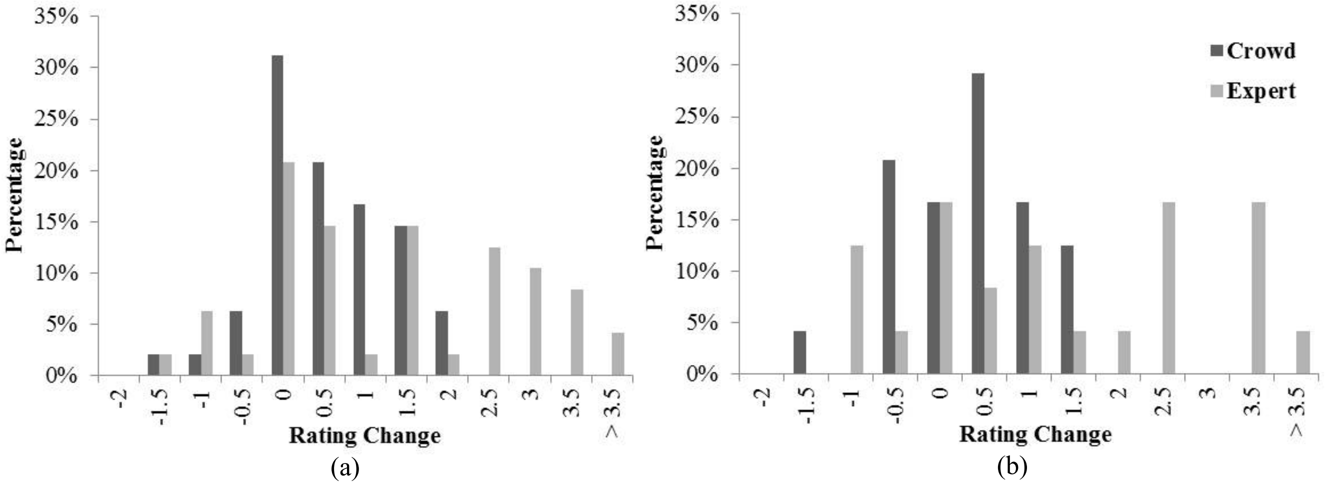


Figure 8. (a) Distribution of the changes in crowd and expert ratings on *Guidelines* across all designs; (b) Distribution of the changes in crowd and expert ratings on *Goals* across all designs.

changes listed in Table 1 only include the changes made based on crowd feedback. Participants may have made additional changes which were driven by their own reflections or the feedback from their peers and instructors. For example, the participant [P9] reported that he did not change the design according to the feedback. However, the color scheme was changed in the revision (see Figure 1).

Theme

Three participants reported creating new visual themes to accommodate crowd responses by adopting different manipulations or choices of an image for the revisions. This type of change often led to other consequential changes. As a result, these participants "*pretty much redid the entire poster*". For example, as shown in Figure 2a, one participant [P3] brightened the main image of the design because her initial design yielded unintended impressions such as "*hard to read*" and "*dark*". After revising, these impressions were no longer present in the feedback (see Figure 2b).

Similarly, the crowd feedback caused another participant [P8] to reconsider the choice and treatment of the images:

> *"The two pictures make the article not too attractive, should put them on another spot or [use] other pictures …"*

> *"The pictures that grab the eye the most are just plastered on like you found them on the internet and pasted them onto a white board with glue, they aren't integrated into the design."*

To address these problems identified from the feedback, the participant decided to replace the images in the poster and reported that the feedback helped him understand how to utilize an image to attract people's attention appropriately:

> *"The emphasis section [First Notice Feedback] helped get a sense of what was attracting the viewers' attention, and some comments helped in getting a better sense of the overall visibility of the information or lack there of."* [P8].

The feedback also helped the participant reconsider the treatment of the images and reorganize visual elements. The feedback on the revision indicates that the first noticed element in the design was successfully changed from the building images to the event (see Figure 3).

Layout

A majority of participants (*n* = 7) revised the layout of their designs based on crowd feedback, such as "*used different grid layout for displaying the text blocks/Sections"* [P8] and "*reorganized the bottom half and simplified the design*" [P2].

One participant [P5] learned that visual elements in the design were overwhelming and caused visual clutter (see Figure 4a). In the revision, the participant repositioned the design content to reduce the visual clutter (see Figure 4b):

> *"Making the poster much less cluttered because, again, a lot of people mentioned that… The goals section [Goals Feedback] is really helpful because it lets me know if I'm doing a good job getting my point across."* [P5]

Another participant [P10] reported that the feedback helped him understand the effects of visual elements and the relationships among them in the design:

> *"I knew which elements worked well with a majority of reviewers, and what elements to specifically keep".* [P10]

As a result, the feedback helped the participant reorganize elements. For instance, she decided to *"change the organization of my speakers and their [information]"*.

Typeface

Many participants (*n* = 6) customized the typeface styles and sizes guided by crowd feedback such as *"the font was too slanted"* [P8] and *"people didn't like the color change of my fonts."* [P6]. As a consequence, participants manipulated their typefaces to address these types of problems:

> *"Changed the type of the texts to make the title standing-out and obvious… It [feedback] helps me to understand which part of my poster stands out the most, and how should I keep making such eye-capturing features in my future works."* [P2]

> *"I changed the spacing of a lot of my text. There were many comments in the impressions section about how cluttered and messy it was, so I took those into consideration when updating my design."* [P10]

Color

According to the survey, participants (*n* = 6) attempted to better match colors to the impressions they intended:

> *"I changed the background color, removing the gray (see Figure 5). People said it was overpowering and looked dreary."*[P7]

> *"I changed the color and gradient effect to make it more natural and realistic."* [P4]

Also, the feedback prompted one participant to change the color scheme of the design in order to increase the contrast between the text and the background. Consequently, the readability of the content was improved:

> *"They find some sections hard to read against a gradient background. So I kept it consistent to one color."*[P6]

DEPTH OF CHANGES (RQ1)

In total, twenty-two changes were described in the surveys and a majority of the changes were on the "significant" side (see Figure 6). The average rating of change depth is 5.2 (σ = 1.1). Most participants (*n* = 9) reported that they changed their designs based on the feedback, and a majority (*n* = 6) agreed the feedback helped them make substantive changes. Participant [P8] anticipated that the first noticed element in the design is the "Avant Garde" event; however, the participant found that the element first noticed by the crowd was the building instead of the event (see Figure 3a). To address this problem, the participant substantively revised the design using different thematic images and reorganized the visual elements. After the changes, the event Avant Garde became the first noticed element in the design (see Figure 3b).

The participants' ratings of their overall depth of changes to a design prompted by crowd feedback positively correlated with the changes in overall rating from the initial to revised designs ($r = 0.56, p < 0.05$). Recall that the overall rating of a design was the mean of the ratings on the guidelines and goals given by the experts for that design (see "Measures"). The overall ratings decreased from the initial to the revised designs for only two participants (P1 and P9), and they made few or no changes based on the crowd feedback. This result indicates that leveraging crowd feedback to iterate on a visual design can lead to improved design outcomes, as measured by the overall rating in our study.

Additionally, all of the feedback types were perceived as helpful to improve designs ($\mu > 4$). For instance, consistent with prior work [36], the *Impressions* feedback was perceived as the most helpful feedback type for participants to improve their designs ($\mu = 5.5, \sigma = 1.0$). The *Guidelines* feedback was the second most helpful type ($\mu = 5.4, \sigma = 1.2$), which was more favorable than in a prior study [36]. This difference is likely due to the emphasis on the teaching and practice of design principles in this particular course.

VALIDITY OF CROWD FEEDBACK (RQ2)

Figure 7 summarizes the non-expert (crowd) and expert overall ratings on the Initial and Revised designs. Table 2 shows the same data, but grouped by the guidelines and the two most commonly stated goals. Non-experts rated the Initial Designs higher than experts (F (1, 244) = 22.04, $p < 0.001$). Non-experts may have lower expectations and therefore were less critical when rating the designs. There was also a small, but statistically significant difference between the crowd's ratings of Initial and Revised designs (paired t-test, $p < 0.01$). This indicates that a non-expert crowd can effectively react to changes made to a design.

To examine how the crowd's reactions compare to experts, we tested the correlation between the rating changes by the crowd and the experts. Several steps were taken to compute the correlation. First, for each design guideline and goal, we computed the average of the ratings for that guideline or goal. For instance, if twenty crowd workers rate how well a design adheres to the *contrast* guideline, then the crowd's *average rating* for this guideline is the average of these twenty ratings. The experts' *average rating* for a guideline is the average of the ratings from the three experts. Second, a *rating change* on a guideline or goal was calculated by subtracting the *average rating* on an initial design from the *average rating* on the revised design for that guideline or goal. The distribution of the rating changes for *Guidelines* and *Goals* are shown in Figures 8a and 8b respectively.

There was a moderate positive correlation between changes in *Guidelines* ratings from the crowd and experts ($r = 0.45$, $p < 0.01$). This indicates that the aggregation of the ratings from non-experts offer a reasonably valid assessment of

	Initial Designs		**Revised Designs**	
	Crowd	Experts	Crowd	Experts
Guidelines (all)	0.88 (0.63)	0.07 (1.68)	1.18 (0.48)	1.41 (0.75)
Proximity	1.07 (0.86)	0.50 (1.88)	1.2 (0.48)	1.75 (0.62)
Alignment	0.81 (0.57)	0.33 (1.66)	1.37 (0.44)	1.58 (0.67)
Repetition	1.03 (0.40)	-0.25 (1.56)	1.19 (0.50)	1.25 (0.81)
Contrast	0.60 (0.57)	-0.29 (1.68)	0.98 (0.46)	1.04 (0.75)
Common goals (all)	1.36 (0.60)	-0.38 (1.67)	1.46 (0.45)	1.00 (1.12)
Grab Attention	1.24 (0.74)	-0.17 (1.81)	1.35 (0.59)	0.88 (1.15)
Provide Details	1.48 (0.40)	-0.58 (1.56)	1.58 (0.23)	1.13 (1.13)

Table 2. The means and (standard deviations) of the crowd and expert ratings of the guidelines and the two most commonly stated goals for the initial and revised designs. The scale of the ratings was from -3 (Worst) to +3 (Best).

Free-from Feedback	Structured Feedback
"Design is excellent looking grand. The font is very nice. I feel pleasant. Very nice"	*"The darker green text on the green. This green instantly made me think of the Statue of Liberty."*
"It is pretty one. The design is marvelous one. The shape of the building is very nice."	*"The words are not aligned properly and also the fonts are very small. The content is not clearly seen"*
"This looks really nice ...and also color combination of black and white color so nice ... and it contains total sessions and date details."	*"Details for each date are clearly below the date and session number, all names/details are grouped together well."*
"This design looks good.it highlights the name of the event, place where it is held, and the dates of the event."	*"It's very colorful and cleat font style. I like this page and this advertisement design also look too good."*
"Design is ok. It should be in light orange and the text should be in navy blue color, small text should be made readable."	*"Light beam is in the background. America is the world in the bounded box... All these elements are perfectly aligned."*

Table 3. A sample of responses from the free-form and structured feedback conditions (n = 1000 in each condition).

improvement in a design based on standard guidelines. We also computed the correlation for each individual guideline. The *contrast* guideline had the highest correlation value ($r = 0.59$, $p < 0.05$), while the *alignment* guideline had the lowest ($r = 0.28$, $p = 0.4$). Consistent with a prior study of crowd feedback [34], non-experts may have experienced different levels of difficulty when applying the different guidelines. A potential solution is to explore scaffolding techniques such as task instructions or examples that can help workers apply the guidelines more effectively [33, 35].

However, there was no statistically significant correlation ($r = 0.37$, $p = 0.07$) between changes in the ratings for *Goals* given by the crowd and experts. The experts thought a revised design better satisfied its goal while the crowd did not agree. The disagreement may be due to different perceptions of how well a goal is satisfied due to different interpretations of the goal or understanding of the context.

The rating changes from the experts were more pronounced than those from the crowd. Figure 8 shows that a large proportion of experts' rating changes was more than 2 units; while most of crowd' rating changes were less than 1. This reflects the fact that experts have higher discriminating ability [14] and thus are more likely to react to changes in a design. Also, a majority of experts' ratings were increased. This result confirms that designers were able to improve their designs in a design iteration based on feedback [12].

	Free-form Feedback			Structured Feedback		
Judgment	Positive	86%	**46%**	Positive	78%	**32%**
	Negative	14%		Negative	22%	
Interpretation	Individual	69%	**42%**	Individual	40%	**64%**
	Relation	31%		Relation	60%	
Suggestion	**12%**			**4%**		

Table 4. Frequencies of the categories of critique discourse found in each form of feedback (*n* = 100 in each condition).

CONTENT ANALYSIS (RQ3)

Our final research question was to compare how generating structured feedback compares to generating free-form feedback on the same twenty designs shown in Figure 1. For each design, we randomly sampled 50 responses generated by the crowd in the free-form and structured feedback conditions (see Table 3). The data set therefore includes 1,000 worker responses in each condition. Each response had on average 22 words ($\sigma = 7.8$). There are no significant differences between free-form and structured conditions with regard to the length of the feedback.

Feedback Genres

Because five types of feedback were specifically generated in the structured condition, we first examined whether free-form feedback spontaneously produces these same types. A random sample of 100 responses from the free-form feedback was coded based on the definition of the feedback types described in [36]. We found only 4 responses that described elements first noticed in a design; 6 responses that commented on design guidelines; and only 8 responses that discussed the intention of a design. Though these types of feedback are known to be desired from non-experts, the free-form condition yielded little content in these areas. Having specific prompts in feedback generation is therefore important for generating the feedback desired by designers.

In order to assess the category differences in free-form and structured feedback, the collected feedback was coded using the nine categories of critique discourse derived in prior work [8]. 100 responses were randomly sampled from both the free-form and structured conditions. The responses were independently coded by two coders (kappa = 0.80) and disagreements were resolved through discussion. Table 4 shows the frequency of occurrence of each category. Several categories of discourse such as *comparison*, *identity invoking*, and *process oriented* did not occur in the collected feedback. One explanation is that these categories were derived from expert feedback [8] and therefore non-experts may not be able to provide this type of insight into users' design approaches and contemporary trends.

Judgment was the most commonly applied feedback category ($\chi^2 = 15.7$, $p < 0.001$; see Table 4). Judgments occurred more often in the free-form feedback (46%) than in the structured feedback (32%). Workers in the free-form condition offered many simplistic judgments, especially related to positive reactions such as "I like it". This result corroborates previous findings that a majority of online feedback does not go beyond simplistic judgments [33, 35].

Feedback coded as *interpretation* occurred when crowd workers tried to make sense of a design and explain their perceptions of it. Interpretation was the most common category in the structured feedback (64%). The structured feedback had higher frequency of interpretation than free-form feedback ($\chi^2 = 5.8$, $p < 0.05$). In the structured feedback condition, workers first provide their overall reaction to a design (e.g. a rating or impression) and *then* explain this reaction. This task separation could help workers pay more attention to the interpretation process.

During the interpretation process, crowd workers in the structured condition more often associated their reactions with elements in a design than workers did in the free-form condition (see Table 4). A worker in the free-form condition attempted to associate his perception with an individual element: *"The tall building gives you the idea of a skyscraper".* In contrast, a worker in the structured condition was more likely to relate several elements in a design: *"Having the text on the building helps establish a spatial relationship in an attractive manner."* One explanation is that workers in the structured condition are exposed to multiple design elements (e.g. design elements offered by the system to help a worker explain the rationale). These specific examples may facilitate the interpretation process and help workers consider the relation among elements in a design.

The *suggestion* category was also significant ($\chi^2 = 7.4$, $p < 0.01$). Free-form feedback offered more suggestions for improvement than the structured feedback (see Table 4). Although crowd workers were not rewarded for offering suggestions in either condition, workers in the free-form condition provided more suggestions. The suggestions were apparently given because there was no constraint imposed by providing a specific prompt. However, the quality of these suggestions needs further investigation.

Topic Diversity

We used LDA to extract the topics from the free-form and structured feedback [4]. LDA is a standard topic modeling method that is often used to discover topics in documents and the words associated with each topic. The method excels at analyzing large amounts of unlabeled documents by clustering words that frequently co-occur. We built LDA topic models from the free-from and structured feedback respectively, treating each response as a document. Since the number of topics affects the interpretability of the extracted topics, we set the number of topics the same in the free-form and structured feedback in order to make a fair comparison. The number of topics was varied between five and twenty, and the differences between the free-form and structured feedback were observed consistently.

Cosine similarity between topic words derived from LDA is used to measure the similarity between topics [1]. We computed the cosine similarity between each pair of topics in the free-form and structured conditions respectively. The similarity values were higher in the free-form condition than in the structured condition (Student's t-test, $p < 0.01$). We observed that many topics in the free-form feedback shared the words such as "design" and "like". It appears that the structured feedback offered a wider range of topics. This indicates that the topics represented in the structured feedback were more diverse than in the free-form feedback. This observation may be best explained by the use of various prompts when generating the structured feedback.

Linguistic Styles

LIWC was adopted to measure linguistic styles for each response received in the free-form and structured feedback conditions. Specifically, each response in the feedback data is measured by percentages of the total words belonging to specific categories. A series of t-test comparisons were carried out between the free-form and structured feedback in terms of the LIWC generated language-use categories. Our null hypothesis is that the free-form and structured feedback are equal in the LIWC features. All 70 categories in LIWC were considered. Due to these 70 simultaneous tests, we allow for a *Bonferroni* correction which adjusts the significance threshold by the total number of statistical tests performed (e.g. $\alpha = 0.05/70 = 0.00014$) [3].

Significant differences were found for affective processes ($p < 0.01$). Crowd workers in the free-form condition used more emotional words than did workers in the structured condition. In many domains, better emotion regulation is a trait of expertise [7, 10]. As described previously, workers in the free-form and structured conditions were unlikely to have design knowledge. Therefore, when non-experts are asked to provide spontaneous judgments of a design, they may rely on affective processes and analyze a design more from an emotional perspective. However, the use of specific prompts and examples in the structured feedback condition may have enabled workers to analyze the design rationally rather than relying solely on affective processes.

The structured feedback had more words per sentence, and long words consisting of six or more letters ($p < 0.01$). A higher ratio of words per sentence and long words is often used as an indicator for better domain-specific working knowledge [18, 19]. One interpretation is that structured workflows facilitate non-experts' ability to interpret a design and communicate their perceptions of it, whereas open-ended prompts do not facilitate non-experts' potential to develop thoughtful responses for creative work.

We found that words related to certainty (e.g. 'always' and 'never') were more common in the structured feedback ($p < 0.01$). Increasing the use of certainty words is an indicator of improved critical thinking [6]. The structured feedback also had a higher rate of articles ($p < 0.01$). Prior work shows that people using articles at a higher rate tend to be more concrete in their thinking [18]. In contrast, free-form feedback was characterized by more spoken categories including more filter and assent words. This result indicates that the free-form feedback was less formal and deliberate.

DISCUSSION AND FUTURE WORK

Our study found that crowd feedback prompted deep changes to the theme and layout of visual designs, and that these and other cosmetic changes led to improved designs as rated by domain experts. The changes were prompted by synthesizing the feedback on *impressions*, *first notice*, *guidelines*, and *goals* generated by the system used in the study. It is hard to tease apart precisely which feedback type or comment prompted each change since design is a reflective and emergent process [28], but each type of feedback was perceived as helpful by the participants.

Though crowd feedback prompted changes that led to improved designs, generating crowd feedback should be considered as a supplement to expert evaluation rather than a replacement. Expert evaluation yields a range and depth of feedback that cannot yet be matched by crowd systems, e.g., analysis of trends, precedents, and design strategies. Our experience to date indicates that crowd feedback is most appropriate when designers need to deliver a clear message to a target audience or need feedback quickly. Future work is needed to explore how to broaden the feedback that can be generated by non-expert crowds.

By comparing non-experts (crowd)' ratings with experts' ratings, we found that non-experts and experts reached better consensus on design guidelines than communicative goals. The result reflects the nature of design feedback, which can be conceived either as a measured judgment governed by universal principles or personal tastes and perceptions [31]. The use of design principles offer a firmer basis for evaluating designs and it is therefore easier to reach agreement. How well a design meets its intended goals is more subjective and varied and therefore more difficult for external evaluators to agree.

The analysis of feedback content in the structured and free-form conditions showed that crowd workers suggest more solutions when prompted to enter free-form feedback. Note that workers were not asked to propose solutions in either condition. Using structured workflows appears to cause workers to respond only as instructed in the task and thus did not propose solutions. We also found many content and stylistic differences between the structured and free-form feedback from the non-expert crowd. In addition to non-experts, it would be interesting to compare how experts perform in both the free-from and structured feedback conditions. Moreover, future work needs to examine how design outcomes are affected by the feedback generated using different formats or crowds with different expertise.

Our study was conducted in context of an entry-level course on visual design. Results from the study point to a benefit of integrating the use of crowd feedback systems into design education. For example, this class of system could be useful in large online courses where students may be unable to obtain sufficient or timely feedback from peers or instructors. In other design courses, crowd feedback may provide a useful supplement to in-class critiques. However, the results of our study may not apply to expert designers since they may have different considerations in their work. For example, experts spend more time evaluating their goals and strategies while novices spend more time being aware of and monitoring design guidelines [26]. Future work is needed to examine how design expertise mediates the use and perception of different types of crowd feedback.

Our study examined how crowd feedback affected a single iteration on a design, but reaching an effective solution often requires multiple iterations. Future work is needed to observe how crowd feedback is utilized throughout an entire design project. Such a study could reveal how crowd feedback benefits different phases of the process. Also, our study focused on one form of visual design, a poster, which is a popular form of visual communication. Our findings should generalize to other forms of visual design such as Web pages, logos, and illustrations. However, an interesting avenue for future work is to determine the types of visual designs for which crowd feedback is most desirable and how often the feedback is requested for these types of designs during the process. It would also be interesting to examine and compare how crowd feedback affects design outcomes in other domains such as architecture.

An assumption of our work is that designers want to utilize crowd feedback to improve their designs. However, there could be unintended uses of crowd feedback. For example, in design education, students may attempt to use favorable crowd feedback to argue for higher scores or to counter the recommendations provided by the instructor or peers. Another consequence is that crowd feedback systems could unintentionally limit creativity in design by shifting too much attention to audience interpretations and suggestions rather than pursuing new and creative directions.

CONCLUSION

Crowd feedback systems offer a new approach for helping designers iterate on their designs, but there has been little research on the effects of these systems. From an empirical study of crowd feedback for visual design, our work has contributed several findings addressing this gap. First, we found that crowd feedback prompted users to make both deep and cosmetic changes to their designs, which led to improved designs. Deep changes related to the theme and layout of a design, while cosmetic changes related to the colors and fonts in a design. Second, we found that a non-expert crowd was able to recognize design improvements based on the use of design guidelines. However, the crowd had lower agreement with experts on the improvements related to communicative goals. Finally, as compared with free-form feedback, the structured feedback produced more interpretative, diverse and critical discourse. Our work shows that crowd feedback systems can be leveraged to help users iterate toward more effective solutions.

REFERENCES

1. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. ACM press, 1999.
2. Bernstein, M.S., Brandt, J., Miller, R.C., & Karger, D.R., Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *UIST*, (2011), 33-42.
3. Bland, J.M., and Douglas G. Altman Multiple Significance Tests: The Bonferroni Method. *British Medical Journal*, *310*, 6973, (1995).
4. Blei, D.M., Ng, A.Y., & Jordan, M.I. Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*, (2003), 993-1022.
5. Blomberg, J.L. and Henderson, A., Reflections on Participatory Design: Lessons from the Trillium Experience. In *CHI*, (1990), 353-360.
6. Carroll, D.W. Patterns of Student Writing in a Critical Thinking Course: A Quantitative Analysis. *Assessing Writing*, *12*, 3, (2007), 213-227.
7. Chaffin, R. and Imreh, G. Practicing Perfection: Piano Performance as Expert Memory. *Psychological Science*, *13*, 4, (2002), 342-349.
8. Dannels, D.P. and Martin, K.N. Critiquing Critiques: A Genre Analysis of Feedback across Novice to Expert Design Studios. *Journal of Business and Technical Communication*, *22*, 2, (2008), 135-159.
9. Dave, B. and Danahy, J. Virtual Study Abroad and Exchange Studio. *Automation in Construction*, *9*, 1, (2000), 57-71.
10. De Groot, A.D. and Groot, A.D.d. *Thought and Choice in Chess*. Walter de Gruyter, 1978.
11. Dow, S.P., Gerber, E., & Wong, A., A Preliminary Study of Using Crowds in the Classroom. In *CHI*, (2013), 227-236.
12. Dow, S.P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D.L., & Klemmer, S.R. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-Efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *17*, 4, (2010).
13. Dutton, T.A. Design and Studio Pedagogy. *Journal of Architectural Education*, *41*, 1, (1987), 16-25.
14. Einhorn, H.J. Expert Judgment: Some Necessary Conditions and an Example. *Journal of Applied Psychology*, *59*, 6, (1974), 562–571.
15. Elkins, J. *Art Critiques: A Guide*. New Academia Publishing, Washington DC, 2012.
16. Feldman, E.B. *Varieties of Visual Experience; Art as Image and Idea*. H.N. Abrams, New York, 1971.
17. Heer, J. and Bostock, M., Crowdsourcing Graphical Perception: Using Mechanical. Turk to Assess Visualization Design. In *CHI*, (2010), 203-212.
18. Kamhi, A.G. and Catts, H.W. *Language and Reading Disabilities*, 2012.
19. Kim, K., Bae, J., Nho, M.-W., & Lee, C.H. How Do Experts and Novices Differ? Relation Versus Attribute and Thinking Versus Feeling in Language Use. *Psychology of Aesthetics, Creativity, and the Arts*, *5*, 4, (2011).
20. Kittur, A., Nickerson, J.V., Bernstein, M.S., Gerber, E.M., Shaw, A., Zimmerman, J., Lease, M., & Horton, J.J., The Future of Crowd Work. In *CSCW*, (2013), 1301-1318.
21. Komarov, S., Reinecke, K., & Gajos, K.Z., Crowdsourcing Performance Evaluations of User Interfaces. In *CHI*, (2013), 207-216.
22. Kulkarni, C. and Klemmer, S. Learning Design Wisdom by Augmenting Physical Studio Critique with Online Self-Assessment. *Stanford University technical report*, (2012).
23. Luther, K., Pavel, A., Wu, W., Tolentino, J.-l., Agrawala, M., Hartmann, B., & Dow., S.P., Crowdcrit: Crowdsourcing and Aggregating Visual Design Critique. In *CSCW*, (2014), 21-24.
24. Mason, W. and Watts, D.J. Financial Incentives and the Performance of Crowds. *SigKDD Explorations Newsletter*, *11*, 2, (2010), 100-108.
25. Park, C.H., Son, K., Lee, J.H., & Bae, S.-H., Crowd Vs. Crowd: Large-Scale Cooperative Design through Open Team Competition. In *CSCW*, (2013), 1275-1284.
26. Perez, R.S. and Emery, C.D. Designer Thinking: How Novices and Experts Think About Instructional Design. *Performance Improvement Quarterly*, *8*, 3, (1995), 80-95.
27. Risatti, H. Art Criticism in Discipline-Based Art Education. *Journal of Aesthetic Education*, *21*, 2, (1987), 217-225.
28. Schön, D.A. *Educating the Reflective Practitioner*. Jossey-Bass, San Francisco, 1987.
29. Strauss, A.L. *Qualitative Analysis for Social Scientists*. Cambridge University Press, 1987.
30. Tohidi, M., Buxton, W., Baecker, R., & Sellen, A., Getting the Right Design and the Design Right. In *CHI*, (2006), 1243-1252.
31. Venturi, L. and Marriott, C. *History of Art Criticism*. Dutton, New York, 1964.
32. Vredenburg, K., Mao, J.-Y., Smith, P.W., & Carey., T., A Survey of User-Centered Design Practice. In *CHI*, (2002), 471-478.
33. Willett, W., Heer, J., & Agrawala, M., Strategies for Crowdsourcing Social Data Analysis. In *CHI*, (2012), 227-236.
34. Xu, A. Designing with Crowds. *University of Illinois at Urbana-Champaign, PhD diss.*, (2014).
35. Xu, A. and Bailey, B.P., What Do You Think? A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *CSCW*, (2012), 295-304.
36. Xu, A., Huang, S.-W., & Bailey, B.P., Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *CSCW*, (2014), 1433-1444.
37. Yu, L., Kittur, A., & Kraut, R.E., Distributed Analogical Idea Generation: Inventing with Crowds. In *CHI*, (2014), 1245-1254.
38. Yu, L. and Nickerson, J.V., Cooks or Cobblers? Crowd Creativity through Combination. In *CHI*, (2011), 1393-1402.
39. Feedbackarmy. http://www.feedbackarmy.com.
40. Fivesecondtest. http://fivesecondtest.com.
41. UITests. http://www.uitests.com.
42. Usabilla. http://www.usabilla.com.