Novice and Expert Sensemaking of Crowdsourced Feedback

EUREKA FOONG, Northwestern University
DARREN GERGLE, Northwestern University
ELIZABETH M. GERBER, Northwestern University

Online feedback exchange (OFE) systems are an increasingly popular way to test concepts with millions of target users before going to market. Yet, we know little about how designers make sense of this abundant feedback. This empirical study investigates how expert and novice designers make sense of feedback in OFE systems. We observed that when feedback conflicted with frames originating from the participant's design knowledge, experts were more likely than novices to question the inconsistency, seeking critical information to expand their understanding of the design goals. Our results suggest that in order for OFE systems to be truly effective, they must be able to support nuances in sensemaking activities of novice and expert users.

CCS Concepts: • **H.5.3 [Group and Organization Interfaces]. Computer-supported cooperative work.**;

Additional Key Words and Phrases: Crowdsourcing; online feedback exchange; crowdsourced feedback; assessment; design; sensemaking; learning; novice; expert

ACM Reference format:
Eureka Foong, Darren Gergle, and Elizabeth M. Gerber. 2017. Novice and Expert Sensemaking of Crowdsourced Feedback. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 45 (November 2017), 18 pages.
https://doi.org/10.1145/3134680

1 INTRODUCTION

45

Ground-breaking designers, Charles and Ray Eames, described design as "a plan for arranging elements to accomplish a particular purpose" [23]. In doing so, designers seek feedback to understand whether they have accomplished their purpose throughout an iterative design process [6].

Until recently, collecting timely, affordable, and authentic feedback was difficult using traditional testing methods such as focus groups and field studies. Increasingly, designers overcome these challenges and supplement traditional testing methods through Online Feedback Exchange (OFE) systems. OFE systems are computer mediated systems that enable individuals to collect feedback from diverse and distributed feedback providers, including online communities, crowdworkers, and potential end users [10]. Design-led companies such as Facebook and Airbnb are among the 34,000 companies who use OFE systems such as *UserTesting.com* [4] to test concepts with more than 1 million potential users. Independent designers also use OFE systems, paying USD$39 a month to access more than 22,000 potential users for feedback on their design concepts through online platforms such as *BetaFamily* [1]. Novice and expert designers can interact with the 486,000 new

This work was generously supported by the Segal Design Institute Design Cluster Fellowship and the National Science Foundation under award 1530837. We would like to thank the paper reviewers, students in the Delta Lab at Northwestern University, Dr. Matt Easterday, Dr. Haoqi Zhang, Emily Wang, and Florence Fu for their extremely useful comments, suggestions, and support.

2573-0142/2017/11-ART45
https://doi.org/10.1145/3134680

members from across the globe who joined graphic design community, *Dribbble* [2], exchanging a total of 841,035 comments in 2016.

Ultimately, successful OFE systems, such as those described above, have the potential to transform the way novice and expert designers conduct user-centered design at scale by providing inexpensive and regular access to designers and potential users. Using OFE systems to complement traditional testing methods can lead to better designs resulting in greater social, environmental, and economic prosperity [25, 26]. But we will fail to realize the potential of computer supported cooperative feedback systems if we do not understand how the systems are being used in practice—specifically how experts and novices make sense of and use the feedback the systems provide [12, 31].

To address this challenge, social computing and CSCW researchers have begun to improve the timeliness, affordability, and quality of feedback obtained through OFE systems. Newly proposed systems prompt high quality critique that is specific, detailed, and positive by providing rubrics of design principles [11, 22, 37] and suggestions for improving feedback [20] to online crowds. Nonetheless, to achieve the vision of transforming user-centered design research and testing, we require a deeper understanding of how these systems can support feedback exchange more holistically, from the time a designer decides to ask for feedback to when the designer revises her work [10]. For example, while it is important for designers to reflect on previous work, many existing systems, including *UserTesting* [4], lack support for comparing design iterations in parallel [10].

Recent research suggests that the relationship between the feedback provided and the designers' subsequent performance is unclear [22, 34]. In one study of an OFE system for novice designers, some participants failed to improve their design work despite reporting the feedback was helpful [34]. In a similar study, novice designers addressed minor, easily-identified feedback, rather than larger synergistic issues with their design [22]. This initial research suggests that the way designers integrate feedback into their understanding of the design task may influence the quality of their revisions.

Yet, we currently lack an understanding of the effectiveness of different approaches to sense-making of feedback. Research suggests that expertise is an important dimension that affects the way individuals approach the design process. Compared to novices, experienced designers spend more time redefining problems, transitioning between design activities, and applying guiding principles to evaluate their work [5, 8]. While we would also expect novice and expert designers to have unique approaches to sensemaking, we currently lack an empirical understanding of these differences.

To address this gap, we conducted a study of 10 novice and 11 expert designers using written feedback from Amazon Mechanical Turk (MTurk) workers. Using the data-frame theory of sense-making [17], we analyzed think-aloud transcripts to understand the frames or structures designers apply to integrate feedback. We examined instances when designers detected inconsistencies with the feedback, as inconsistencies provide important insight into the frames that are applied to a task [17]. Our analysis reveals five frames that designers use to make sense of feedback and five ways they typically react to information inconsistent with these frames.

The results suggest that novice and expert designers require different forms of support in order to successfully integrate feedback. Specifically, design experts are more likely than novices to consider the process of collecting feedback and seek information to expand their understanding of the design challenge. We contribute implications for designing OFE systems that better support novice and expert designers in using crowdsourced feedback. Before CSCW researchers continue to invest significant effort in developing crowdwork [16], it is critical to consider how diverse users will make use of the products of crowdwork.

2 BACKGROUND

Online feedback exchange is a method for requesting and receiving information from distributed users online to improve the quality of creative work, including design [10]. Examples of systems that support OFE include online communities of creatives (e.g., *PhotoSIG* [33], *Dribbble* [2]), platforms for sharing creative work with stakeholders and end users (e.g., *Mural* [3]), and usability testing websites that employ crowdworkers (e.g., *UserTesting* [4]). OFE systems have gained recent attention in the CSCW community because of their ability to help creatives collect constructive critique quickly from as many as 500 people in as little as 24 hours [35]. While OFE systems' ability to aggregate feedback efficiently in this way provides a cost-effective solution to populations without regular access to face-to-face feedback, many systems struggle to maintain high quality across all critique [14], a problem consistently identified by crowdsourcing researchers [15].

2.1 Effectiveness and Limitations of Crowdsourced Feedback

The CSCW community has made significant progress in understanding the characteristics of effective online feedback. In addition to specificity and criticality [37], positively framed feedback from anonymous feedback providers increases perceptions of the helpfulness of critique [14, 24]. The motivation of feedback providers also affects the length of critique, as well as the number of suggestions and process-oriented feedback provided [36]. Researchers find that system factors can influence the type of feedback provided. For example, numeric ratings encourage online peer reviewers to provide explanatory, but lower quality feedback [13]. Researchers have used these characteristics to develop a natural language model that gives feedback providers suggestions for improving their critique [20].

Although these interventions have improved the perceived helpfulness of feedback [35], some designers still struggle to make improvements based on crowd feedback. Researchers observed that participants using the *CrowdCrit* feedback tool addressed easily-identified and popular issues, rather than deeper issues with the proposed design [22]. An empirical understanding of how designers integrate feedback into their understanding of the design task is critical for CSCW researchers committed to improving the effectiveness of OFE systems and, more broadly, crowd-supported cooperative work. Our study addresses this gap by comparing the sensemaking processes of expert and novice designers.

2.2 Models of Sensemaking

Sensemaking is the process of imposing order on a disorganized set of data to guide future actions that refine one's understanding of a situation [32]. It is a way for experts to test hypotheses with others and develop a shared understanding of a situation. For example, in a medical setting, organizational theorist Karl Weick and colleagues observed that inconsistencies in patient data trigger the sensemaking process in nurses, who then examine and label relevant symptoms and communicate hypotheses to other nurses [32]. Likewise, in a military setting, Pirolli and Card observed how intelligence experts made sense of data to decide military plans [27]. Based on these observations, they proposed several sensemaking activities that systems should support, including determining relevant data, visualizing connections between data, and prompting users to consider alternative hypotheses. Underlying these models are the processes of recognizing and refining frames to explain a set of disorganized data.

According to Klein and colleagues' data-frame theory, frames (e.g., scripts, maps) are cognitive structures that explain the relationships between events or data [18]. When inconsistent data is introduced, frames can either be preserved or refined (i.e., reframed) [17]. For example, if a nurse observes a symptom that is inconsistent with the current diagnosis, he or she can either continue to preserve the diagnosis or propose new explanations about the patient's situation. Therefore, inconsistent data can reveal characteristics of an existing frame.

While many theories of sensemaking are primarily based on expert models of the process, Klein and colleagues find that sensemaking is influenced by domain expertise [18]. Sensemaking studies of intelligence analysts suggest that expertise influences the sophistication of frames and responses to inconsistent data [18, 19]. Participants in one study were asked to read a scenario about transporting college students by bus in a conflict-ridden area of Kosovo [19]. Information operations experts were more likely than novices to investigate reasons behind the drop in ridership in this scenario and make more accurate conclusions about the situation. Similarly, differences in sensemaking activities by novice and expert designers could influence their ability to make effective design revisions.

2.3 Differences Between Expert and Novice Sensemaking in Design

Like sensemaking, design is an iterative process of generating and evaluating solutions to refine one's understanding of a design challenge [8, 21, 29]. Scholars have described design as a reflective practice, drawing on Schon's description of reflection-in-action [8]. Experienced designers spend more time redefining the design challenge, generating and evaluating changes to the design, and transitioning between different design activities [5]. In addition, exemplary designers use strong "first principles," such as visual contrast, as frames to guide their design work [8]. These studies suggest that expert designers may have better frames and strategies for solving issues that arise with their design and suggests that less experienced designers may need different support to effectively use feedback in OFE systems.

While data-frame theory has been empirically investigated in medical and military settings, we do not yet know if the theory holds in a design context where multiple satisfactory outcomes exist [29] and it can be difficult to assess the "accuracy" of one's understanding of the design challenge. In this study, we close this empirical gap by providing insight into the sensemaking processes involved in using OFE systems. In addition to improving the design of these systems, these new insights contribute a deeper understanding of the psychological process of sensemaking in the design context.

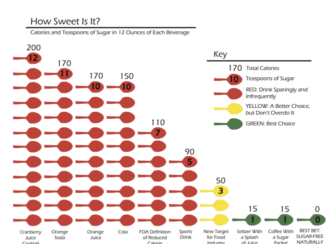
Our high-level research question was: how are expert and novice designers similar and different in their sensemaking processes when integrating crowdsourced feedback? We asked the following specific research questions:

- **RQ1: What types of frames do novice and expert designers use to explain the relationship between feedback comments?** Drawing from theories of sensemaking [17], we hypothesized that expert designers are more likely than novices to use their rich design knowledge as a frame to relate feedback comments.
- **RQ2: How do novice and expert designers respond to feedback that is inconsistent with their frames?** In line with prior work [19], we hypothesized that expert designers are more likely than novices to refine their frames by seeking additional information about inconsistent feedback.

3 METHOD

To better understand the needs of diverse designers who use OFE systems, we aimed to investigate the sensemaking processes of expert and novice designers. To compare designers' thought processes, we followed prior work on sensemaking (e.g.,[19]), crowdsourced feedback [35], and expert-novice behavior in peer production communities [31] to conduct a controlled study using the think-aloud protocol and interviews.

Table 1. Posters and scenarios used in the study.

Title	Scenario	Poster
Poster A	This designer wants to create an attractive poster to recruit college students for a research study. The poster will be posted in buildings on campus and must contain information that will help participants decide whether and how to participate. You will soon see feedback collected from five college students.	
Poster B	This designer wants to create an attractive poster that can be posted in doctor's offices and online to warn parents of choking hazards in the home. You will soon see feedback collected from five parents with young children at home.	
Poster C	This designer wants to create an attractive poster that can be shared online through social media to attract visitors to a website about healthy eating. The poster is about different types of drinks and the amount of sugar in these drinks. You will soon see feedback collected from five adults who frequently use social media.	

3.1 Participants

We recruited 21 participants (10 novices; 10 female; age range 18-64) from a mid-sized private university and a local design consultancy. The novice designers had no prior professional or academic experience in graphic or communication design and had limited experience with graphic design software, such as Adobe Photoshop. The experts all had designed web interfaces and/or managed graphic designers, and their experience ranged from three to 25 years. The majority had taken at least three courses on graphic or communication design and used graphic design programs more than once a week. All regularly received and integrated user feedback into their work. We offered all participants USD$30 in return for participating; however, four expert designers declined the compensation.

3.2 Stimuli

We presented participants with crowdsourced feedback provided by 15 MTurk Workers on three different poster designs (Table 1). To ensure quality crowdsourced feedback, we followed best practices when soliciting feedback on posters from Mechanical Turk workers [11, 16, 37]. Specifically, we included example critique statements, asked workers to answer objectively verifiable questions

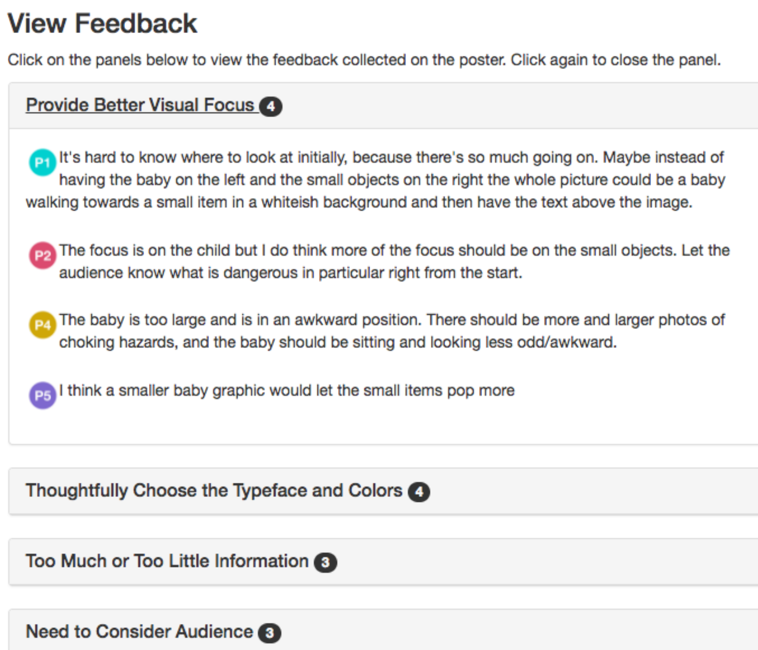


Fig. 1. A screenshot of the feedback interface shown to participants. Each poster was shown to the left of the feedback panels, with all panels closed at the beginning of the study.

about the design, and recommended that workers provide positive and specific feedback. The task took workers an average of 25 minutes and 47 seconds to complete. Workers were compensated at a rate of USD$9/hour.

In line with prior work [22], we asked five MTurk workers to critique each poster (total: 15 MTurk workers). This allowed us to collect a number of feedback comments comparable to previous studies (M = 24.7) [22]. To replicate Luther and colleagues' findings [22], workers provided feedback using the same rubric of eight design principle categories in *CrowdCrit*. We then designed a website to display the crowdsourced feedback in the rubric alongside the poster and design scenario (Figure 1). In this display, we presented the participants with the comments beside a thumbnail representing each individual feedback provider (e.g., P1, P2, P3, P4, P5).

3.3 Procedure

Following prior sensemaking studies [19], we engaged each participant in a think aloud protocol. To familiarize the participants with the think aloud protocol, we asked each participant to search for a pair of rain boots on Google.com for one minute, speaking aloud as they completed the task. Once confident they understood the protocol, we asked them to imagine themselves as design consultants and to read the design scenario for each poster for one minute (see Table 1). Participants were told that the feedback was from the target audience of each poster. Following this, they were asked to think aloud while viewing the feedback and preparing to improve each poster. We asked participants to spend 12 minutes per poster and counterbalanced the sequence for showing these posters to participants. In addition to recording their voice and actions on the screen, we provided participants with paper and a text box on the computer to take notes if they desired to do so.

After following the think aloud protocol, participants rated their confidence in their understanding (1: not very confident, 10: very confident) and the helpfulness of the feedback across all posters (1: not very helpful, 10: very helpful). Then, we asked them questions about their strategies for applying feedback, any additional information they would want to ask of feedback providers, and if they noticed inconsistent feedback between feedback providers. Finally, participants completed a survey about their design experience and usage of graphic design software. The entire procedure was tested and refined in a pilot study one month earlier.

3.4 Data Analysis and Coding

All interview and interaction sessions were recorded and transcribed. Subsequently, we wrote memos after each participant [30]. During the memo writing process we noticed that participants frequently talked about different frames they used to make sense of the data and the instances when feedback conflicted. We also noted that participants reacted to inconsistencies by questioning the data, proposing new changes to the design, simulating changes proposed by feedback providers, comparing feedback, or ignoring the inconsistency. We began a process of selective coding of the transcripts in which we flagged each event when participants noticed conflicting feedback. After identifying each of these events, we clustered the events into conceptual categories based on 1) the source of the inconsistency (i.e., type of frame) and 2) response to the inconsistency. Simultaneously, we researched pertinent literature to understand existing theory and uncover related phenomena. Initial data analysis began after 10 interviews.

There were a total of 351 events across 21 participants. Reviewing the transcripts independently, two coders identified all events in which participants detected an inconsistency in the feedback. Next, the two coders independently assigned codes to a sample of 20% of the events for each participant ($N = 76$) to establish inter-rater reliability [7] (Kappa, frames = 0.75; Kappa, responses to inconsistencies = 0.71). This was in line with prior work on crowdsourced feedback where 10% of the data was sampled for inter-rater reliability [11, 37]. To further improve reliability, our coders discussed disagreements on the response codes. One coder consistently assigned codes even though the statements were not specific enough to be responses (e.g., "maybe this should be addressed later on something else"). Once we revised our definition, we achieved higher inter-rater reliability (Kappa, responses to inconsistencies = 0.84). One of the coders then completed coding on the remaining events in the dataset.

4 RESULTS

We first determined all instances when participants noticed inconsistencies between feedback or disagreed with feedback. Both novices ($M = 7.15$, $SD = 0.81$) and experts ($M = 8.09$, $SD = 1.70$) reported a high level of perceived understanding, and there was no significant difference between the groups ($p = 0.13$). Novices ($M = 7.75$, $SD = 1.50$) and experts ($M = 6.32$, $SD = 2.15$) also found the feedback moderately helpful ($p = 0.10$). For clarity, we refer to study participants using the following numbering scheme: P1001, P1002, etc.

4.1 RQ 1: Types of Frames

Participants used five types of frames to evaluate and integrate feedback: 1) their personal preferences, 2) their knowledge of the design goals and target audience, 3) their understanding of the feedback process, 4) previous feedback comments, and 5) categories in the feedback rubric (see Table 2 and Table 3 for examples and percentages of events by poster). To compare novice and expert designers, we analyzed the tendency for each group to use different frames. A chi-square test of independence showed a significant relationship between expertise and the frequency of types of frames used, χ^2 $(4, N = 351) = 75.99$, $p < .0001$. This relationship was only significant when the

Table 2. By noticing inconsistencies in the comments, participants demonstrated that they were using several types of frames to make sense of feedback.

		Novice Example	Expert Example
Types of Frames	Personal Preferences	”If the whole background were white, it would just be blanking and annoying. I like the blue gradient, it’s reminiscent of babies.” (P1019)	”Handles I don’t mind... I can recognize that these are spoons pretty quickly. But if you got rid of the handles I don’t think it’d be clear.” (P1021)
	Knowledge of Design Goals	"A wheelchair might limit our audience if we’re looking for people with other disabilities as well." (P1006)	"...this starts to touch on the overall goal of this...Because if this poster answers all of the questions you have, then it’s not doing the goal of moving people to the website" (P1016)
	Process of Collecting Feedback	"I think it would be more interesting to hear from more people who have disabilities...what they thought of the poster." (P1005)	"I would like to show them the poster and ask ’What are the top three objects babies choke on?’ It feels like we asked them the wrong questions." (P1022)
	Previous Feedback Comments	"We have three people who say we need to change the color and three people who are fine with color." (P1008)	"Potentially conflicting information about spoons and text and whether or not that works...I guess they’re just saying that having numbers and spoons together work better to communicate."(P1009)
	Rubric Categories	"The layout’s a little off... Why didn’t they include that in the Create a More Sensible Layout feedback? I don’t know." (P1018)	"I’m surprised they didn’t put that in the Layout part." (P1025)

process frame was included in the analyses; when we removed those instances, the chi-square test was no longer statistically significant, χ^2 (3, N = 300) = 4.07, p > 0.1. This suggests that novice and expert designers used similar types of frames, but experts were more likely consider the process of collecting feedback (see Figure 2). In the following paragraphs, we provide examples of each type of frame used by novices and experts.

In our study, novices and experts used similar types of frames to explain the connections between feedback. To our surprise, both experts (33.9% of events) and novices (53.9% of events) most frequently used personal preferences to frame and evaluate the usefulness of feedback. Participants in both groups admitted to looking for comments that reflected their own opinions, rather than the design’s goals or well-established design principles. As one expert with more than 30 years of visual communication experience explained:

> "My instinct is to consider... what validates what I already believe. When I first looked at the poster, I had certain things that I thought needed to be done.” (P1023, Poster A)

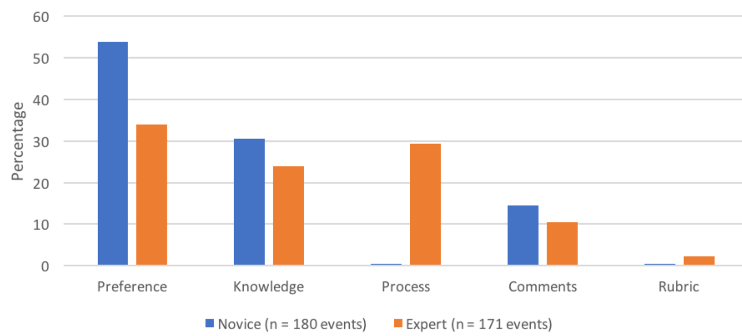


Fig. 2. Types of frames as a percentage of events coded for each group of participants. Experts were significantly more likely to apply their understanding of the feedback process during sensemaking compared to novices.

A novice designer who has never attended a design course before agrees with the expert's approach:

> "I think I have the kind of poster that I want in mind. I'm just like looking at the feedback to see whether my hypothesis is true or not." (P1008, Poster A)

Participants applied the Personal Preferences frame when they disagreed with a feedback provider's evaluation or suggestions for improvement. For instance, one novice designer (P1015) who uses graphic design programs less than once a month disagreed with the suggestion to move the title "Prevent Choking!" to the top of Poster B. Instead, he deferred to his own opinion that the title was the focus of the poster and that the placement was appropriate.

> "I actually think it's a good thing to have 'Prevent Choking!' in the middle. It's very much the focus. I don't necessarily know that what's on the top is the focus all the time, so I think this is a smart placement." (P1015)

Expert designers similarly acknowledged that they were processing the comments through personal "biases" and "filters." When looking at Poster B, one expert designer (P1012) expressed that it was harmful for her to disagree with feedback because she was not the intended audience of the poster. Nevertheless, she relied on her initial perceptions to point out a lack in feedback about font alignment, which none of the comments directly addressed:

> "There's alignment issues with the font, just like before, the font wrapping around the baby image makes it hard to read. Nobody said that." (P1012)

Novices (30.6% of events) and experts (24.0% of events) also noticed when feedback was inconsistent with their objective knowledge of the design's goals and the target audience. One feedback comment that consistently went against this frame was a suggestion to add an image of a wheelchair to Poster A. Several novice and expert designers disagreed with this suggestion, noting that disabilities are not merely physical. Furthermore, one novice designer said that the goal of the poster was not specifically to recruit the physically disabled:

> "P5's opinion is interesting but it doesn't seem like to me that the poster is specifically for physically handicapped people." (P1024)

Apart from applying the Knowledge of Design Goals frame, expert participants were significantly more likely than novices to notice flaws in the feedback collection process and the characteristics of the feedback providers. The Process of Collecting Feedback frame represented 29.2% of the

Table 3. Percentages of coded events by poster and participant group. Applying personal preferences and not responding to inconsistent feedback were most prevalent for both novice and expert designers across all posters.

		Poster A		Poster B		Poster C	
		Novice	Expert	Novice	Expert	Novice	Expert
Types of Frames	Preference	53.1	41.9	59.7	36.5	46.3	27.6
	Knowledge	32.7	23.3	22.1	27.0	40.7	22.4
	Process	2.0	23.3	0	17.3	0	40.8
	Comments	12.2	9.3	16.9	17.3	13.0	6.6
	Rubric	0	2.3	1.3	1.9	0	2.6
Responses to Inconsistencies	Seek Information	10.2	18.6	3.9	32.7	5.6	26.3
	New Goal	14.3	18.6	11.7	7.7	18.5	7.9
	Simulate Design Change	6.1	2.3	9.1	0	3.7	4.0
	Compare Comments	0	0	3.9	7.7	1.9	4.0
	None	69.4	60.5	71.4	51.9	70.4	57.9

inconsistencies detected by experts, but was only used once by one novice participant. Some experts were sensitive to feedback providers' lack of expertise. On Poster C, expert designer (P1010) who has managed a wide range of designers, commented on the characteristics of the feedback and the feedback providers. This suggests that he was building up an idea of how the feedback was collected in order to make sense of the data:

> "Well I don't think this user understood what a visual hierarchy was because they just said they find it helpful." (P1010)

The expert designer then described why the feedback itself was unhelpful:

> "I think people are just saying 'Yes, we should have visuals and text.' I don't think they gave us anything useful in terms of how we balance those." (P1010)

Finally, the expert designer drew conclusions about the experience of the feedback provider:

> "They're just not that visually critical or sensitive." (P1010)

Other expert designers seemed to have an idea of how feedback should have been collected. Applying her professional experience, P1022 disagreed with the way we showed feedback providers the poster designs. Instead, she said she would have shown multiple versions of the designs in parallel to the feedback providers. P1022 repeatedly said that feedback providers were being asked to comment on specific visual details of the posters, rather than on what the posters communicated to them:

> "If this is how our design firm is operating, we need to seriously reconsider our process. We put posters in front of users and asked them to comment on the graphic design, instead of their likelihood to get the message...I would like to show them the poster and ask 'What are the top three objects babies choke on?' It feels like we asked them the wrong questions." (P1022)

All experts mentioned preferring feedback on how the posters were experienced, rather than feedback on how to specifically improve the posters. One expert who designs visual assets in her career wanted less directive feedback because the suggested solution may need to be modified:

> "Usually when I give graphic design feedback, some things are very hard and fast. 'You should change this color because it's impossible to read.' But most of the time, I try to frame it as, 'Maybe try this...'... I think it's hard for me personally to say, 'Oh, if you put the text below the image, it will be perfect.' That's not how I work when I do this stuff. I put it on there, and just shift... I'm not convinced that if you put the text on the other side of the picture, everything will be perfect." (P1009)

Aside from that, experts (10.5% of events) and novices (14.4% of events) also noticed inconsistencies between comments made by the same and by different feedback providers, suggesting that they use previous comments to make sense of feedback. One novice designer (P1006) applying the Previous Comments frame noticed that only one feedback provider had made a negative comment about the font in Poster B. P1006 chose to agree with the majority of the feedback providers, who stated that the font was appropriate. In response to a suggestion to remove information from the graph in Poster C, one expert noted that this was inconsistent with previous feedback:

> "So it seems like that really could be addressed more with simplifying the design rather than taking out information because it looks like the rest of the people in the group seemed to like the amount of information that they're getting." (P1027)

Experts (2.3% of events) and novices (0.6% of events) infrequently noticed when feedback was inconsistent with the titles of categories in the rubric. The Rubric Categories frame was different from the Process of Collecting Feedback frame because participants were expecting feedback to align with the titles of the categories. They did not necessarily have expectations for how the feedback should be collected. This often occurred when the title suggested there was negative feedback when there was also positive feedback in the category. One novice designer describes her desire to to see consistent feedback in one category:

> "If 'Provide Better Visual Focus' is the topic.... I would put a comment there if I felt [the designer] needed to provide a better visual focus. But sometimes people would say, 'Oh the focus is great.' But others would say, 'Oh it looks a little cluttered.' I can't tell what's going on. Having those comments all together makes it hard to judge what the overall thing is." (P1005)

4.2 RQ 2: Responses to Inconsistent Feedback

When participants encountered inconsistent feedback, they either 1) ignored the feedback (None), 2) sought additional information (Seek Information), 3) proposed design goals or changes (Propose Design Goal), 4) simulated suggested design changes (Simulate Suggested Changes), or 5) compared comments to one another (Compare Comments) (see Table 3 for percentages of events by poster, Table 4 for examples, and Figure 3). As in RQ1, we compared the tendency for novices and experts to respond in these different ways. A chi-square test of independence showed a significant relationship between expertise and the type of response to inconsistent feedback, χ^2 (4, N = 351) = 32.42, p <.0001. The relationship was only significant when we included events when participants sought additional information; when we removed those events, the chi-square test was no longer statistically significant, χ^2 (3, N = 295) = 4.22, p > 0.1. This suggests that expert designers (26.32% of events) were more likely to respond to inconsistent feedback by seeking additional information compared to novice designers (6.11% of events). In the following paragraphs, we provide additional examples of each type of response.

In the majority of cases, experts (56.7% of events) and novices (70.6%) did not explicitly respond to inconsistent feedback, instead preserving the original frame they had used to evaluate the feedback. For example, after reading a suggestion on Poster B to use a white background, one novice designer said he liked the blue background and did not engage further with the comment:

Table 4. When participants responded to inconsistent feedback, they did so by seeking information, proposing new design goals, simulating suggested changes, and comparing comments.

		Novice Example	Expert Example
Responses to Inconsistent Feedback	Seek Information	"I think that highlighting something as red, yellow, or green, might depend on your age... I would need to find that information out before I publish this ad." (P1007)	"If I had one question to ask the participants, I would want to know. How do you feel? Are you afraid?" (P1010)
	Propose Design Goal	"So if I were to do this, I would make a bar graph. I would get rid of the spoons...." (P1017)	"I would be interested in... possibly switching to a vertical layout so that you could have the image... the headline, image, and all of the body copy." (P1027)
	Simulate Suggested Changes	"It is hard to visualize this feedback in my head and how it would map to the poster itself. It requires a big overhaul I think" (P1024)	"But as a poster size, that much red and yellow is not particularly attractive." (P1021)
	Compare Comments	"P4 talks about the fine print and the phone numbers and that goes back to P2's aspect about typeface and colors being larger. P4 doesn't think that a phone number is even relevant and the website is good enough." (P1006)	"It seems 1 and 5 have pretty neutral, not very meaty feedback, whereas I would say 3 and 4 have more design expertise or actually care about voicing their opinion more." (P1026)
	None	"I don't think yellow would be a good idea though." (P1015)	"So this is one of those categories where people give praise. It is not helpful at all. I would discount that." (P1023)

> "If the whole background was white, it would be blank and annoying. I like the blue, it's reminiscent of babies." (P1019)

Another expert designer disagreed with a suggestion that viewers of Poster C would count all of the spoons on the poster:

> "That's interesting feedback, but I don't think everybody would think that so... I'll maybe wrap that point into an outlier for me. I think making people count all of these and keep track can be a pain. I hate doing that myself." (P1021)

As mentioned earlier, expert designers were more likely to respond to inconsistent feedback by seeking additional information. For example, one expert disagreed with feedback providers that the main image in Poster B should be made smaller. Her response was to seek information and plan an A/B test with two versions of the poster to understand why the image was problematic:

> "...everybody seemed to think that the baby should be smaller. I didn't really agree with that, but I might try and do an A/B test or two concepts for the next round. One that has a smaller baby, just to try it out, and more choking items, and then one that has

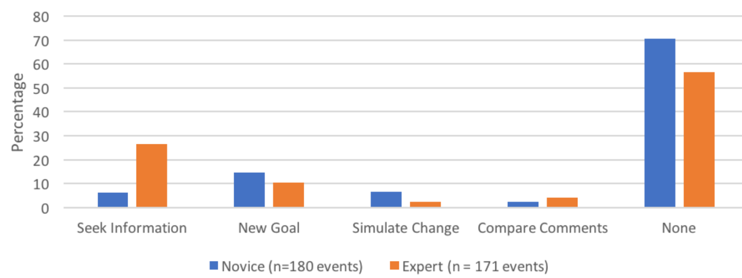


Fig. 3. Strategies for responding to inconsistent feedback as a percentage of events coded for each group of participants. Experts were significantly more likely to respond to inconsistent feedback by seeking additional information compared to novices.

> basically the same size baby, and a single, but kind of more threatening choking item." (P1003)

Participants used the Seek Information response to reflect on the goals of the poster, consider new opportunities, and gather more critical information about the design challenge. For example, a feedback provider on Poster C commented that "the graph is great for learning information, especially for those who are visual learners over auditory learners." While some participants felt this suggestion contradicted the goals of the poster and ignored the feedback, one expert designer responded by questioning whether a poster was necessary:

> "Hmm...Interesting to think about. Since it's going to be seen on social media, should this even be a poster? Should it be a video?" (P1022)

In another case, one expert designer (P1025) noticed a discrepancy between comments on Poster B, in which one feedback provider wanted the contact information of the government agency to be more prominent while another felt this information was not important. This prompted her to reevaluate the goals of the poster:

> "I would want to learn more, or really evaluate what our goal is here. So do we want to warn parents of choking hazards? Maybe a next step would be to confirm with the design team the goal and corresponding call to action" (P1025)

Some novice designers also used the Seek Information response when faced with inconsistent feedback. They considered alternative explanations behind the feedback provider's motives. For example, P1005 disagreed with the feedback that there was too much white space in Poster B. Instead, P1005 considered that the feedback provider may have only raised the issue because they had been asked about it:

> "This person is worried about empty space. Generally, I think empty space can help make things more clear. I kind of wonder if that's something that they only thought of because they were asked to look at this poster in more depth." (P1005)

Another way participants would respond to inconsistent feedback was to Propose Design Goals. Both novices (14.4% of events) and experts (10.5% of events) reacted to inconsistent feedback by planning new design changes. For example, one novice (P1008) disagreed with a suggestion on Poster A to use a photo of a person in a wheelchair. Instead, she proposed that the picture be removed altogether and that the poster be made more colorful to address the problem. Another expert designer (P1016) disagreed with feedback that Poster A contained too much information, proposing that some of the information be made clearer instead:

> "I think the disability could be more clear and then, again, targeting the college student, what's the time commitment... and how is it going to fit in their schedule..." (P1016)

In a few cases, experts (2.3% of events) and novices (6.7% of events) would Simulate Suggested Changes and imagine how the feedback would influence the effectiveness of the poster. One expert designer read a comment on Poster C about emphasizing drinks to consume, rather than avoid. Although the designer disagreed with the suggestion to "rethink the whole thing," the designer simulated these changes by imagining how the chart would look if the order of the drinks were rearranged:

> "I don't know if you need to redo the whole thing...I'm not sure there's something here, but if you took the inverse of this that goes from something bad and low... over to the right, which is good. I don't know if you can do anything with that, but I think it's worth exploring, but we could try the inverse of the chart." (P1010)

Although infrequent, novices (2.2% of events) and experts (4.1% of events) would also Compare Comments between feedback providers when faced with inconsistent feedback. For example, one novice designer used her physical notes to tally the number of feedback providers who commented on the font color in Poster B:

> "They all have like different opinions on this. Some like the colors, but some think there are problems with the colors...one person noted the font, but nobody else really talked about it." (P1006)

Finally, we ran a Cochran-Mantel-Haenszel test to understand connections between types of frames and responses to inconsistent feedback. Regardless of design expertise, there was a significant general association between type of frame used and response to inconsistent feedback, χ^2 (16, N = 351) = 35.6, $p < 0.01$. This suggests that for both novices and experts, the type of response given was contingent on the type of frame used. This significant association was driven primarily by the Personal Preferences frame. When all instances of this frame were removed, the general association was no longer significant, χ^2 (12, N= 196) = 17.5, $p > 0.1$. In 69.7% of all instances when the Personal Preferences frame is used, novices and experts did not respond to the inconsistency. This suggests that for both novices and experts, feedback that is inconsistent with one's preferences is primarily ignored.

5 DISCUSSION

Using data-frame theory as a lens, we identified five types of frames designers use to make sense of crowdsourced feedback: 1) their personal preferences, 2) knowledge of the design goals and target audience, 3) the process of collecting feedback, 4) previous feedback comments, and 5) categories in the feedback rubric. When feedback was inconsistent with these frames, designers responded by 1) ignoring the inconsistency, 2) seeking additional information, 3) proposing new goals or changes to the design, 4) simulating suggested changes to the design, and 5) comparing feedback. We discovered two aspects in which novice and expert designers differed in their sensemaking: expert designers were more likely to use the process of collecting feedback as a frame and to seek information about inconsistent feedback.

Our interviews provide early insight into the mechanisms that underlie these differences. Experts attributed their focus on the feedback process to experiences in their own or their students' projects. By asking users how they experienced a design, rather than what they would change about it, experts found they could propose more creative and effective solutions. Additionally, working with professional clients reminded experts that feedback providers often do not mean what they say. Therefore, experts developed adaptive responses when they disagreed with feedback providers, such as asking follow up questions and showing additional design variations. These findings begin to explain why experts use different types of frames and responses during sensemaking. While previous research in human-computer interaction and information retrieval focused exclusively on expert models of sensemaking [27, 28], we show that it is just as important to consider how novices make sense of information. We found that as designers develop expertise, they may benefit from different forms of crowdsourced feedback. In comparison to novices, expert designers more frequently considered the process of collecting feedback when making sense of crowd feedback. They wanted comments that were less directive and more indicative of how the design was experienced. In line with this, prior work suggests that design instructors provide more advanced students with open-ended critique rather than directive suggestions [9]. Future studies should consider how designers of various expertise levels benefit from different types of crowd feedback.

Our findings indicate that Online Feedback Exchange (OFE) systems should provide additional support to novices to engage with inconsistent feedback. In line with data-frame theory [18], expert designers questioned inconsistent feedback more frequently than novice designers, enabling them to reflect on their design goals, consider new issues, and construct plans to collect more information (e.g., A/B tests). This demonstrates how feedback that is dismissed by one designer can become a source of useful information by another when the feedback prompts further questioning. Nonetheless, participants as a whole did not respond to the majority of inconsistent feedback. They had a tendency to preserve their original frame, particularly when feedback was inconsistent with their personal preferences. This tendency to preserve frames has yet to be explored in depth in sensemaking research [18], but could indicate that knowing *when* to dismiss feedback may be an acquired skill. Future work should continue to explore the potential benefits and drawbacks of preserving frames during sensemaking.

While we predicted that expert designers would rely on their design knowledge to make sense of feedback, both experts *and* novices relied heavily on their preferences and knowledge of the design goals during sensemaking. These findings are probable given that design has been described as a highly intuitive process[8]. This begins to explain why some novice designers in prior studies failed to examine deeper connections between crowdsourced feedback [22]. Despite having less expertise, novices were just as likely as experts to use their personal preferences and knowledge of the design goals to evaluate feedback. However, it is currently unclear if these self-evaluations accurately reflect the issues within a design. We will empirically investigate the impact of relying on personal preferences and design knowledge on subsequent revisions in future work.

Our findings have broader implications for CSCW and crowd-supported cooperative work systems. Prior work has focused on increasing the efficiency of sensemaking systems based on expert models of the sensemaking process [27, 28]. In order to be truly effective, however, these systems must consider the expertise and unique experiences of their users. In this study, expertise influenced how users integrated and responded to inconsistent feedback, suggesting that users will find different types of information important. Moreover, we raise concerns about novice designers using personal preferences to guide their response to inconsistent feedback. In short, we contribute an expanded understanding of expert *and* novice sensemaking in systems that support iterative design. Further, we reinforce the critical importance of understanding the experience of multiple user groups in CSCW systems.

5.1 Design Implications

Building on prior CSCW design strategies for systems that serve both experts and novices [31], we offer the following design strategies. In comparison to novice designers, expert designers were keen to improve the process for collecting feedback. Therefore, OFE systems can support more advanced designers by providing alternative forms of crowd feedback. Instead of encouraging crowdworkers to provide specific suggestions, OFE systems should encourage them to provide detailed explanations about their understanding or experience of the design. Systems should also support advanced designers by providing additional contextual information about how the feedback was collected.

Our results suggest that designers use several types of frames to impose order on crowd feedback and often notice inconsistencies between feedback and these frames. Seeking information about these inconsistencies can reveal critical information about the design challenge. Therefore, OFE systems can support designers by helping them quickly discover these inconsistencies. In this study, for example, the system could have allowed designers to set aside inconsistent comments into new categories or automatically generated statistics about the attitudes of feedback providers toward certain elements in the design (e.g., the baby in Poster B). By helping designers aggregate inconsistent feedback, OFE systems can encourage them to ask critical questions to advance their project.

5.2 Limitations and Future Work

One of the limitations of our study is the scope of the system we presented to participants. In reality, OFE systems present feedback in different ways across multiple contexts [10]. Because our goal was to investigate differences in approaches to sensemaking, we chose to replicate the feedback interface from recent work as closely as possible [22, 34]. Nevertheless, as Pirolli and Card [27] suggest, different presentation styles may influence the cost of sensemaking. For example, we speculate that OFE systems that automatically aggregate feedback into "word clouds" might prompt different frames compared to systems that display feedback as unordered lists. The current study provides a valuable starting point and methodology for examining these nuances in OFE systems.

The design of our study also limited our analysis to types of frames that novice and expert designers use, rather than the complexity of those frames. We chose this analysis for two reasons. First, data-frame theory does not provide an adequate operational definition of frame complexity [19]. Second, participants relied heavily on their subjective preferences and knowledge of design goals during sensemaking. Hence, analyzing the complexity and connection between frames would be highly unreliable with independent coders. Future research might address this limitation by providing a more constrained study task to participants with fewer feedback comments or posters.

While many designers must revise designs given to them, many conceive of and revise their own work. To compare the sensemaking process across participants, we chose three third-party designs. Weick's theory of organizational sensemaking [32] predicts that frames related to one's identity can be especially difficult to modify, even in the face of inconsistent evidence. Therefore, an important area for future research is understanding whether these findings are supported when designers receive and apply feedback on their own work.

Besides that, to compare poster designs, we asked participants to only *prepare* to revise each poster instead of revising the posters in full. As a result, we lack a detailed understanding of how designers use feedback during the revision process. The interactions designers have with feedback and their emerging solutions may have an additional influence on the quality of their design work. Hence, future research should examine designers' interactions with crowdsourced feedback to iteratively improve their design.

Another avenue of research is to examine how designers make sense of inconsistent feedback from expert rather than non-expert crowds. Several expert participants in this study disliked being told how to improve the posters. One participant noted that he could have gotten more constructive feedback from two or three design experts. Moreover, prior research suggests that designers perceive feedback from novice designers as less helpful than feedback from more experienced designers [37]. Negative perceptions of feedback from novice designers could certainly influence the process of sensemaking, and future work should investigate whether designers will react differently to inconsistent feedback from professional designers.

6 CONCLUSION

Our results apply data-frame theory to the domain of crowdsourced design feedback. We found that novice and expert designers used five types of frames to integrate feedback into their understanding of the design challenge. All designers noticed when feedback was inconsistent with their personal preferences, knowledge of the design goals, previous feedback comments, and categories in the feedback rubric, but only experts were sensitive to the process of collecting feedback. Expert designers were more likely than novices to welcome inconsistencies as opportunities to seek information about the design challenge. In order to be effective, OFE systems must support these nuances in sensemaking. More broadly, when developing effective crowd-powered CSCW systems, developers must design for the unique experiences of users with different expertise. With this perspective, we may realize the full potential of CSCW systems to help all people across their lifespan.

ACKNOWLEDGMENTS

This work was generously supported by Northwestern's Design Research Cluster Fellowship and the National Science Foundation under award 1530837. We would like to thank the paper reviewers, students in the Delta Lab at Northwestern University, Dr. Matt Easterday, Dr. Haoqi Zhang, Emily Wang, and Florence Fu for their extremely useful comments, suggestions, and support.

REFERENCES

[1] [n. d.]. BetaFamily. ([n. d.]). Retrieved July 7, 2017 from www.betafamily.com
[2] [n. d.]. Dribbble. ([n. d.]). Retrieved July 7, 2017 from www.dribbble.com
[3] [n. d.]. MURAL. ([n. d.]). Retrieved July 7, 2017 from www.mural.ly
[4] [n. d.]. UserTesting. ([n. d.]). Retrieved July 7, 2017 from www.usertesting.com
[5] Cynthia J Atman, Justin R Chimka, Karen M Bursic, and Heather L Nachtmann. 1999. A comparison of freshman and senior engineering design processes. *Design Studies* 20, 2 (1999), 131–152.
[6] Sara L Beckman and Michael Barry. 2007. Innovation as a learning process: Embedding design thinking. *California Management Review* 50, 1 (Oct. 2007), 25–56.
[7] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (1968), 213–220.
[8] Nigel Cross. 2004. Expertise in design: An overview. *Design Studies* 25, 5 (2004), 427–441.
[9] Deanna P Dannels and Kelly Norris Martin. 2008. Critiquing critiques: A genre analysis of feedback across novice to expert design studios. *Journal of Business and Technical Communication* 22, 2 (April 2008), 135–159.
[10] Eureka Foong, Steven P Dow, Brian P Bailey, and Elizabeth M Gerber. 2017. Online feedback exchange: A framework for understanding the socio-psychological factors. In *Proceedings of the ACM Conference on Human Factors in Computing Systems.*
[11] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM Conference on Creativity and Cognition.*
[12] Jonathan Grudin. 1988. Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 85–93.
[13] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 458–469.
[14] Julie Hui, Amos Glenn, Rachel Jue, Elizabeth Gerber, and Steven Dow. 2015. Using anonymity and communal efforts to improve quality of crowdsourced feedback. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing.*
[15] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. ACM Press, New York, New York, USA, 64.

[16] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the 26th Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, New York, USA, 453–456.
[17] Gary Klein, B Moon, and R R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE intelligent systems* 21, 5 (2006), 88–92.
[18] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah Peluso. 2007. A data-frame theory of sensemaking. In *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*.
[19] Gary Klein, W Seick, D Peluso, and J Smith. 2007. *FOCUS: A model of sensemaking*. Technical Report.
[20] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.
[21] Bryan Lawson. 2005. *How Designers Think: The Design Process Demystified*. Elsevier, Oxford, England.
[22] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM Press, New York, New York, USA, 473–485.
[23] John Neuhart, Charles Eames, Ray Eames, and Marilyn Neuhart. 1989. Eames Design: The Work of the Office of Charles and Ray Eames. Harry N. Abrams.
[24] Duyen T Nguyen, Thomas R Garncarz, Felicia Ng, Laura Dabbish, and Steven Dow. 2016. Fruitful Feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*.
[25] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.
[26] Victor Papanek and R Buckminster Fuller. 1972. *Design for the Real World*. Thames and Hudson, London.
[27] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*.
[28] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '93*. ACM Press, New York, New York, USA, 269–276.
[29] Herbert A Simon. 1988. The science of design: Creating the artificial. *Design Issues* 4, 1/2 (1988), 67.
[30] James P Spradley. 1980. *Participant Observation*. Wadsworth Publishing Company.
[31] Jennifer Thom-Santelli, Dan Cosley, and Geri Gay. 2010. What do you know? Experts, novices and territoriality in collaborative systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1685–1694.
[32] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the Process of Sensemaking. *Organization Science* 16, 4 (Aug. 2005), 409–421.
[33] Anbang Xu and Brian P Bailey. 2012. What do you think?: A case study of benefit, expectation, and interaction in a large online critique community. In *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, New York, New York, USA, 295–304.
[34] Anbang Xu, Shih-Wen Huang, and Brian P Bailey. 2014. Voyant: Generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social computing*. 1433–1444.
[35] Anbang Xu, Huaming Rao, Steven P Dow, and Brian P Bailey. 2015. A classroom study of using crowd feedback in the iterative design process. In *the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM Press, New York, New York, USA, 1637–1648.
[36] Yu-Chun Grace Yen, Steven Dow, Elizabeth Gerber, and Brian P Bailey. 2016. Social network, web forum, or task market? Comparing different crowd genres for design feedback exchange. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM Press, 773–784.
[37] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Vennix, Steven P Dow, and Björn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on the perceived value of crowdsourced design critique. In *the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

Received April 2017; revised September 2017; accepted November 2017