TEXT AND VISUAL ANNOTATION TOOLS FOR SCALABLE DESIGN FEEDBACK

BY

KIRILL MANGUTOV

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Brian P. Bailey

ABSTRACT

Designers who wish to solicit feedback online have access to a variety of tools. Yet when selecting one tool over another for feedback collection, there is little empirical evidence to guide a designer's decision. We conducted an online study (N=360) where participants provided design feedback using two representative classes of feedback collection interfaces: spatial and non-spatial. For each interface, we also manipulated access to history feedback. Our results showed that the presence of history introduced a fixation effect where providers entered feedback that was more similar to the feedback they reviewed. Providers in the non-spatial condition entered feedback that was 24% longer than the spatial condition; whereas providers in the spatial condition left more *investigation* feedback. There was no difference in specificity between conditions. Results suggest that the more important choice designers must make is not the class of tool they use but whether history feedback is included.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Designers iterate towards solutions that better connect with the target audience by collecting and addressing user feedback [1]. Soliciting design feedback is more scalable, available, and affordable when done online [2, 3]. But soliciting feedback online is possible using a variety of different tools, and this choice could affect the quality of the feedback received.

One way of conceptually organizing these tools is by the feedback provider interface. A spatial interface is one where before entering feedback the provider must first visually mark a location on the design. Spatially marking the design requires the provider to visually search the design and focus their attention on specific elements [3, 4]. This approach may encourage only surface-level feedback. Examples of this class of tool are Adobe Acrobat and Redpen.io [5, 6]. Tools in the non-spatial class implement a text-centric open ended interface. Providers enter feedback into a prominent text area. The text area input requires providers to reference the design as a whole. Referencing the design as a whole may elicit more conceptual feedback. Reddit and Dribble are examples of this class of tool [7, 8].

In this work, spatial and non-spatial classes of tools are compared. This comparison is important because these are the two more widespread classes of online design feedback tool. Prior research has found that even small differences in the interface can significantly alter the provider's behavior. Providers enter feedback that is more diverse when a task is split into multiple shorter sub-tasks [9]. Higher quality feedback is generated when scaffolding is employed [10, 11].

Another consideration is how revealing the feedback from prior providers (or *history*) within a feedback collection tool influences subsequent providers. Granting providers access to history feedback may enhance creativity and encourage novel ideas [12-14]. But it may also reduce feedback diversity by causing a fixation effect [15, 16]. Fixation has primarily been studied in context of giving examples during idea generation and synthesis [14]. Our work will study the possibility of this effect for writing design feedback – a more analytic task.

In this paper, we solicit design feedback using four interface conditions. Interface conditions included two classes of popular real-world feedback collection tools, spatial and non-spatial. For each of these two interfaces, we manipulated the presence of history. Collected feedback characteristics are studied in each condition.

We recruited participants (N=360) to provide feedback on three categories of designs across all conditions. In each condition, the goals of the design and the design image were reviewed by the participant. They then entered feedback using the assigned interface. If history was present, providers had the opportunity to review the feedback left by prior participants. After entering the feedback, participants completed a self-assessment survey.

For the feedback, we measured length, similarity to reviewed history, and specificity, and we analyzed the frequencies of specific categories of feedback. We also measured task completion time and analyzed effort and usefulness from the self-assessment ratings.

Our main findings include that presence of history introduces a fixation effect. Feedback generated in the non-spatial interface was 24% longer and had more stop words. There was no difference in feedback specificity between conditions. We also found that classes of interface produce different categories of feedback. Feedback of the *investigation* category was more likely to be generated in the spatial interface. Our results show that the decision to use a spatial or non-spatial interface would be tied to whether the designer wants longer feedback or more *investigation* feedback. The more important decision is whether one would choose a tool which includes history in either of these interfaces as our results show it introduces a fixation effect. This fixation effect could lead to less diverse content, yet diversity of perspective is one of the reasons designers would choose to use an online feedback tool. We believe our results will contribute to helping designers know how choice of tool influences the received feedback.

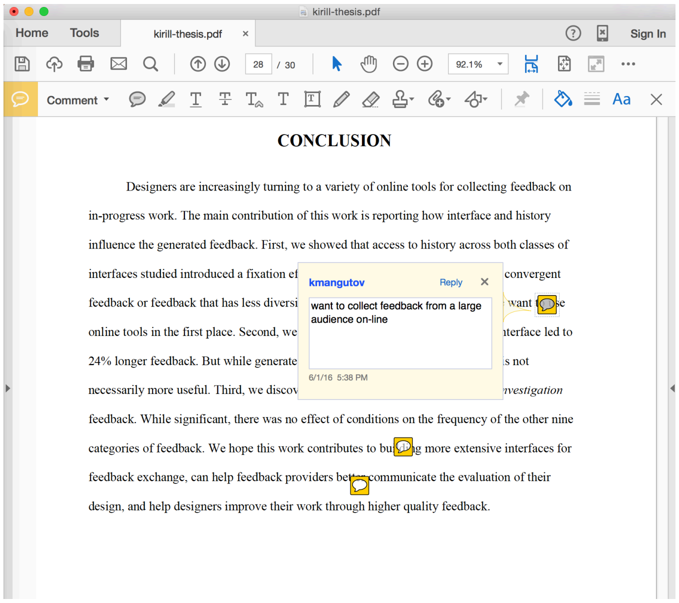
CHAPTER 2

RELATED WORK

We build on two main areas of related work: feedback collection tools, and studies of crowd feedback systems.

2.1 Feedback Collection Tools

There are at least four classes of online tools for collecting design feedback and conducting peer review. Spatial annotation tools require the feedback provider to first select a region of the design to enter feedback. Requiring feedback providers to visually search for and mark features can focus their attention on specific elements [17]. But this focus may also introduce a fixation effect, causing an inability to see new ways of problem solving [14]. Adobe Acrobat and Redpen.io, shown in Figure 2.1 and Figure 2.2 respectively, implement this class of tool [5, 6].



CONCLUSION

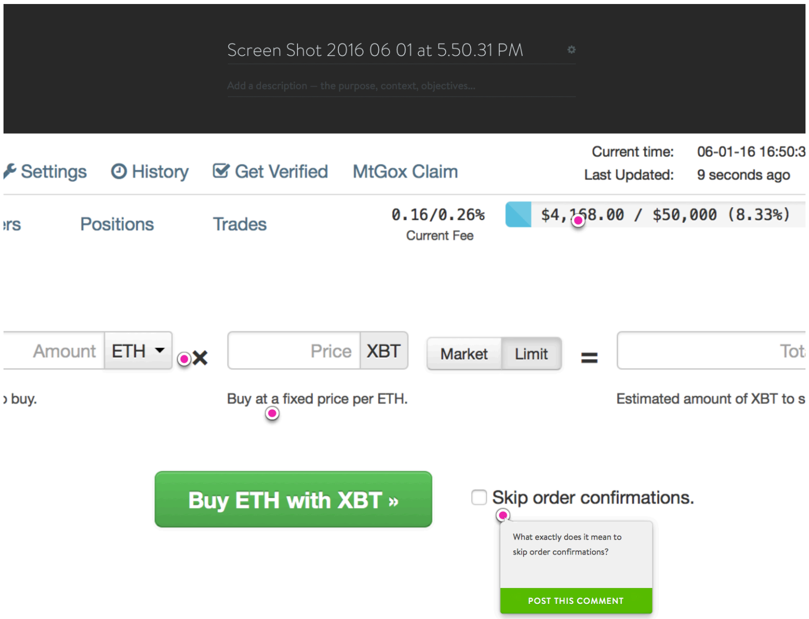
Designers are increasingly turning to a variety of online tools for collecting feedback on in-progress work. The main contribution of this work is reporting how interface and history influence the generated feedback. First, we showed that access to history across both classes of
interfaces studied introduced a fixation ef convergent
feedback or feedback that has less diversi want t se
online tools in the first place. Second, we nterface led to
24% longer feedback. But while generate s not
necessarily more useful. Third, we discov *nvestigation*
feedback. While significant, there was no effect of conditions on the frequency of the other nine categories of feedback. We hope this work contributes to bu g more extensive interfaces for feedback exchange, can help feedback providers be communicate the evaluation of their design, and help designers improve their work through higher quality feedback.

kmangutov Reply

want to collect feedback from a large audience on-line

6/1/16 5:38 PM

Figure 2.1. The Adobe Acrobat feedback provider interface. Before entering feedback, the provider must select a region on the document. Selecting a region requires the provider to visually scan the design.

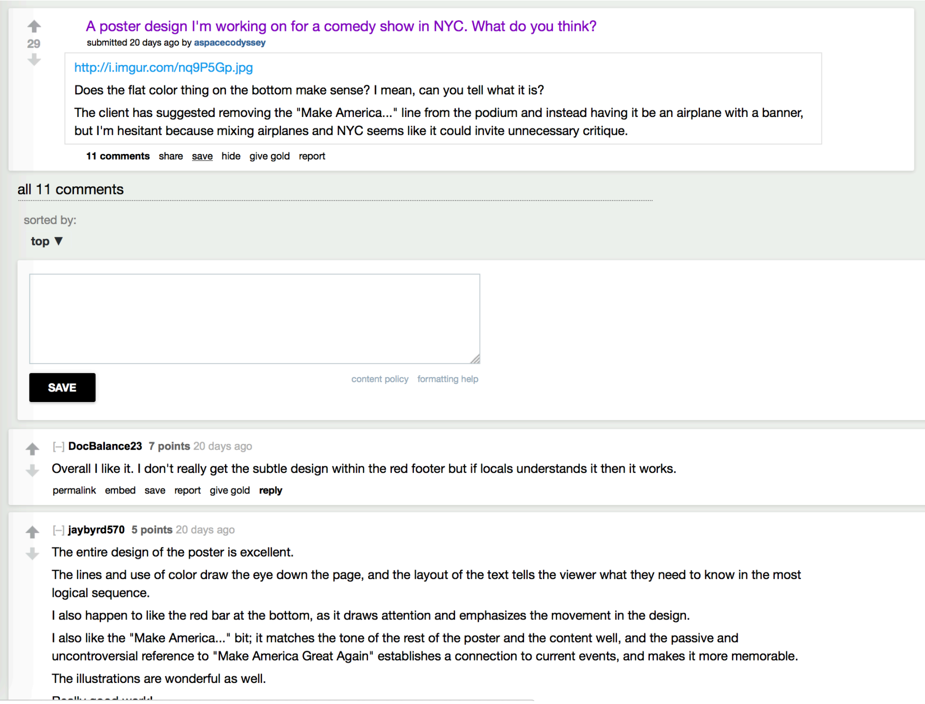


What exactly does it mean to skip order confirmations?

Figure 2.2. The Redpen.io feedback provider interface. The design image is prominently displayed. Requiring the provider to visually scan the design may cause him to focus on specific elements in the design.

Non-spatial tools present an image of the design and have providers enter feedback into a textbox. A prominent textbox may encourage the feedback provider to generate longer, more conceptual feedback because providers must reference the design as a whole. But this less actively engaging interface paradigm could reduce the diversity of generated feedback [8]. Reddit and Dribble, shown in Figure 2.3 and 2.4 respectively, implement this class of tool [7, 8].

A more general class of peer review tool is multi-modal. Multi-modal tools track pen hovering movements in tandem with voice and digitizer writing. These tools have been shown to be preferred to in-person meetings by students [18]. But the linear and irreversible nature of voice makes the commenting task more stressful since providers had to think and speak at the same time. An implementation of this tool is RichReview++, shown in Figure 2.5.



A poster design I'm working on for a comedy show in NYC. What do you think?
submitted 20 days ago by aspacecodyssey

29

http://i.imgur.com/nq9P5Gp.jpg

Does the flat color thing on the bottom make sense? I mean, can you tell what it is?

The client has suggested removing the "Make America..." line from the podium and instead having it be an airplane with a banner, but I'm hesitant because mixing airplanes and NYC seems like it could invite unnecessary critique.

11 comments share save hide give gold report
all 11 comments
sorted by:
top ▼
SAVE
content policy formatting help

[–] DocBalance23 7 points 20 days ago

Overall I like it. I don't really get the subtle design within the red footer but if locals understands it then it works.

[–] jaybyrd570 5 points 20 days ago

The entire design of the poster is excellent.

The lines and use of color draw the eye down the page, and the layout of the text tells the viewer what they need to know in the most logical sequence.

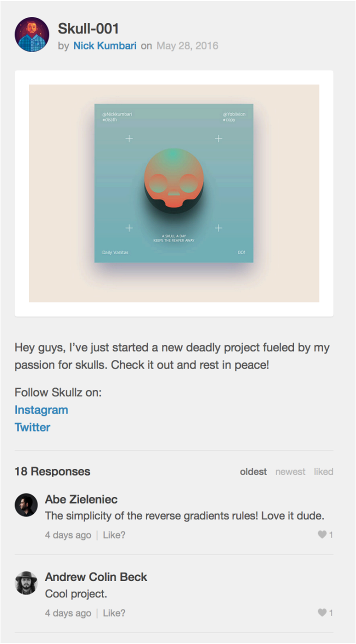
I also happen to like the red bar at the bottom, as it draws attention and emphasizes the movement in the design.

I also like the "Make America..." bit; it matches the tone of the rest of the poster and the content well, and the passive and uncontroversial reference to "Make America Great Again" establishes a connection to current events, and makes it more memorable.

The illustrations are wonderful as well.

Really good work!

Figure 2.3. The Reddit feedback provider interface. The provider enters feedback into a prominent text area. It is not necessary for the provider to first select a region on the design.



Skull-001
by Nick Kumbari on May 28, 2016

Hey guys, I've just started a new deadly project fueled by my passion for skulls. Check it out and rest in peace!

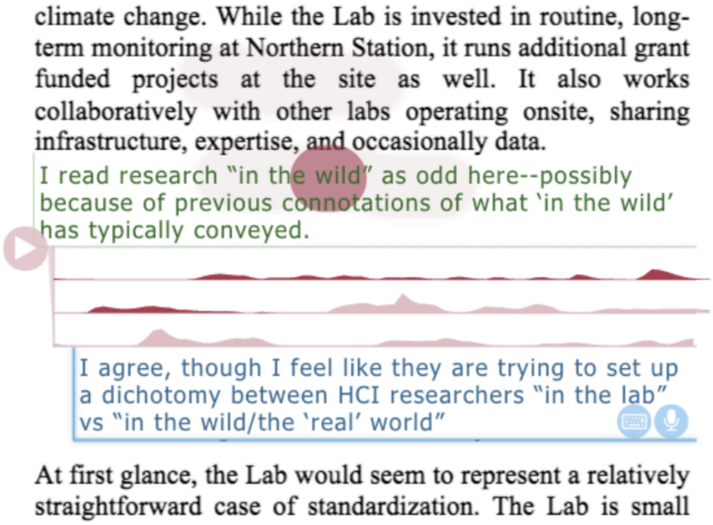
Follow Skullz on:
Instagram
Twitter

18 Responses

Abe Zieleniec
The simplicity of the reverse gradients rules! Love it dude.
4 days ago | Like? ♥ 1

Andrew Colin Beck
Cool project.
4 days ago | Like? ♥ 1

Figure 2.4. The Dribble online feedback tool. The provider leaves a text comment. Leaving a text comment may influence the provider to reference the design as a whole.



climate change. While the Lab is invested in routine, long-term monitoring at Northern Station, it runs additional grant funded projects at the site as well. It also works collaboratively with other labs operating onsite, sharing infrastructure, expertise, and occasionally data.

I read research "in the wild" as odd here--possibly because of previous connotations of what 'in the wild' has typically conveyed.

I agree, though I feel like they are trying to set up a dichotomy between HCI researchers "in the lab" vs "in the wild/the 'real' world"

At first glance, the Lab would seem to represent a relatively straightforward case of standardization. The Lab is small

Figure 2.5. The RichReview++ interface. The provider enters multi-modal annotations containing text, voice, and gestures. Borrowed from Figure 1 of "RichReview++."

Visual tool providers use an image browser to compile their feedback. Image feedback is especially useful for communicating first impressions [19]. But limiting providers to using images may make them unable to convey their ideas. Moodsource is an example of this class of tool, shown in Figure 2.6.

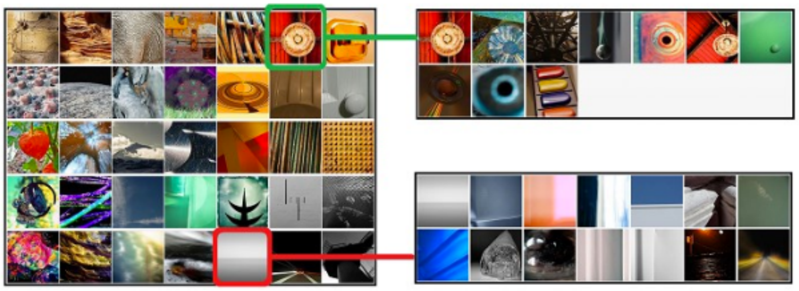


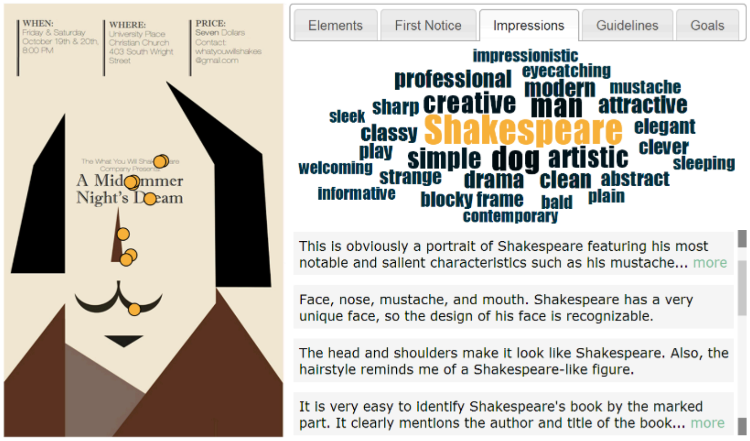
Figure 2.6. The Moodsource image browser. The provider used the image browser to compile a visual summarization of the design. Moodsource allowed intuitive visual communication between crowds and designers. Borrowed from Figure 3 of "Moodsource."

Our work targeted the two more widespread classes of online design feedback tool, spatial and non-spatial. We also studied these classes with the presence of history. Presence of history feedback could increase provider creativity [12, 14]. But it could also introduce a fixation effect [14]. Examples of tools that show history are Redpen.io and Reddit [4, 7].

There has been little prior work on how different classes of feedback collection tool or how the presence of history influences generated feedback. Our work addresses this gap.

2.2 Studies of Crowd Feedback Systems

Crowd feedback systems has been shown to lead to improvements in designs [2]. One such system, Voyant, provided designers with up to five categories of feedback [3]. After providers selected the categories desired, the system would create sub-tasks and submit them to an online labor market. Individual task outcomes were aggregated and presented to the user. Feedback was presented using a bi-directional interaction technique which linked overviews of content and annotations on the design. Designers found Voyant, shown in Figure 2.7, useful in analyzing relations between perception of a design and the visual elements within [3].



This is obviously a portrait of Shakespeare featuring his most notable and salient characteristics such as his mustache...

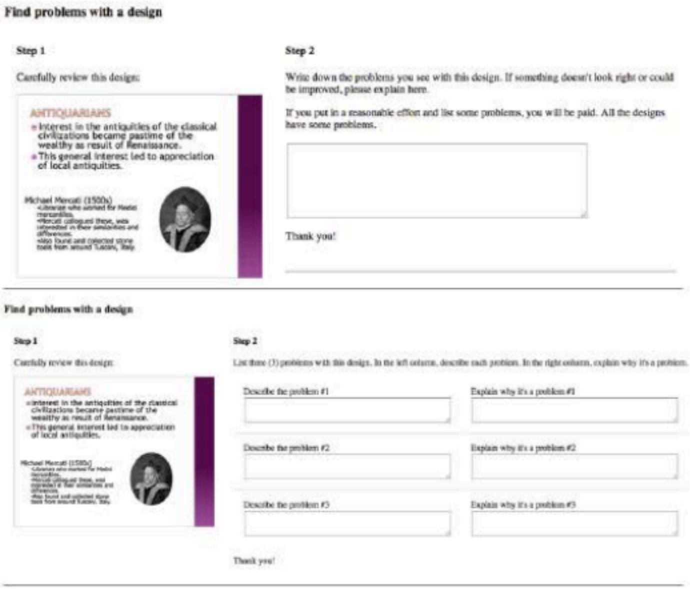
Face, nose, mustache, and mouth. Shakespeare has a very unique face, so the design of his face is recognizable.

The head and shoulders make it look like Shakespeare. Also, the hairstyle reminds me of a Shakespeare-like figure.

It is very easy to identify Shakespeare's book by the marked part. It clearly mentions the author and title of the book...

Figure 2.7. The Voyant crowd feedback system. The system created sub-tasks and aggregated outcomes for the user. Voyant focused on the experience of the feedback receiver. Borrowed from Figure 1 of “Voyant.”

CrowdCrit produced high-quality crowd critique through scaffolding [10]. A series of seventy pre-authored critique statements were available to providers to compile their feedback. After feedback collection was completed, the response distribution was then shown to the user, revealing the highest priority issues. This information was generally found helpful by designers. The CrowdCrit system is shown in Figure 2.8.



Find problems with a design

Step 1

Carefully review this design:

ANTIQUARIANS
- Interest in the antiquities of the classical civilizations became pastime of the wealthy as result of Renaissance.
- This general interest led to appreciation of local antiquities.

Michael Mercati (1500s)
- Librarian who worked for Medici mercantiles.
- Mercati catalogued these, was interested in their similarities and differences.
- Also found and collected stone tools from around Tuscany, Italy.

Step 2

Write down the problems you see with this design. If something doesn't look right or could be improved, please explain here.

If you put in a reasonable effort and list some problems, you will be paid. All the designs have some problems.

Thank you!

Find problems with a design

Step 1

Carefully review this design:

ANTIQUARIANS
- Interest in the antiquities of the classical civilizations became pastime of the wealthy as result of Renaissance.
- This general interest led to appreciation of local antiquities.

Michael Mercati (1500s)
- Librarian who worked for Medici mercantiles.
- Mercati catalogued these, was interested in their similarities and differences.
- Also found and collected stone tools from around Tuscany, Italy.

Step 2

List three (3) problems with this design. In the left column, describe each problem. In the right column, explain why it's a problem.

Describe the problem #1

Explain why it's a problem #1

Describe the problem #2

Explain why it's a problem #2

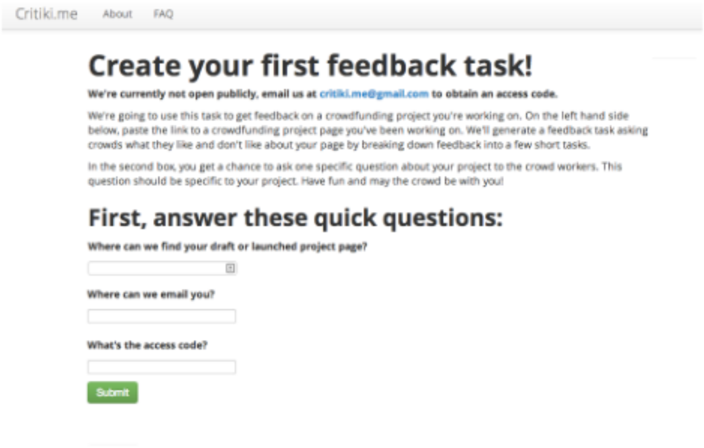
Describe the problem #3

Explain why it's a problem #3

Thank you!

Figure 2.8. The CrowdCrit crowd feedback system. A scaffolded interface was shown to enhance generated feedback quality. Borrowed from Figure 2 of "Structuring, Aggregating, and Evaluating Crowdsourced Design Critique."

Critiki introduced a system that simplified the process of collecting feedback [11]. The system created, distributed, and aggregated crowdsourced design critique for crowdfunding pages. Effective critique was encouraged through scaffolding. Examples of high-quality critiquing points to assist providers in composing useful critique accompanied worker prompts. The Critiki system is shown in Figure 2.9. Our work differs by comparing two classes of user interfaces for the feedback provider.



Create your first feedback task!

We're currently not open publicly, email us at critiki.me@gmail.com to obtain an access code.

We're going to use this task to get feedback on a crowdfunding project you're working on. On the left hand side below, paste the link to a crowdfunding project page you've been working on. We'll generate a feedback task asking crowds what they like and don't like about your page by breaking down feedback into a few short tasks.

In the second box, you get a chance to ask one specific question about your project to the crowd workers. This question should be specific to your project. Have fun and may the crowd be with you!

First, answer these quick questions:

Where can we find your draft or launched project page?

Where can we email you?

What's the access code?

Figure 2.9. The Critiki crowd feedback system. The system managed the process of generating design feedback tasks and aggregating results.

CHAPTER 3

METHODOLOGY

This thesis compares how generated feedback is influenced by two classes of feedback Interface (spatial and non-spatial) and History (absent and present). We seek to answer whether these conditions cause the provided feedback to be more specific or general, if they influence the likelihood of generating a certain category of feedback, and if the presence of history introduces a fixation effect.

These questions are not exhaustive but are intended to give designers a better sense as to how their choice of feedback collection tool will influence feedback received online. The results may also create awareness among system developers as to how their implementation choices influence the feedback exchange.

3.1 Experimental Design

To answer these questions, we conducted a full-factorial, between-subjects experiment. The factors were Interface (Non-spatial vs. Spatial) x History (Absent vs. Present) x Design Category (Poster vs. Webpage vs. Web Interface), giving a 2x2x3 design.

3.1.1 Participants

Mechanical Turk was used to recruit feedback providers (N=360). Providers were required to have successfully completed at least 50 tasks and to have a task approval rate greater than 95%. In total, 206 males and 154 females age eighteen and over participated. Based upon a pilot study, the payment was set at $0.50 per task to reflect current US minimum wage.

3.1.2 Designs

We chose three designs, selected to span a broad range of visual domains, to be familiar to a general audience, and to warrant design improvements. The selected designs included a poster advertising a university dance event, the home page of a community college (http://parkland.edu), and a web-based payment application (https://venmo.com/). Explicit permission from the creator of the first design was obtained and the two remaining designs were public domain. The designs are shown in Figure 3.1.

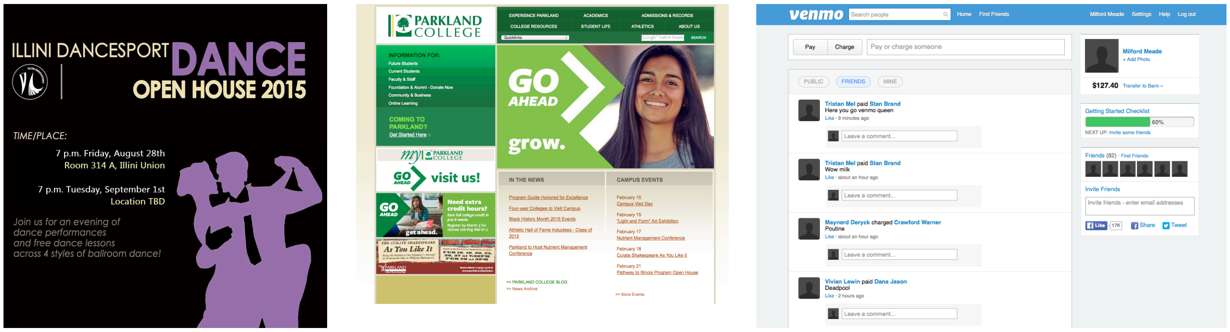
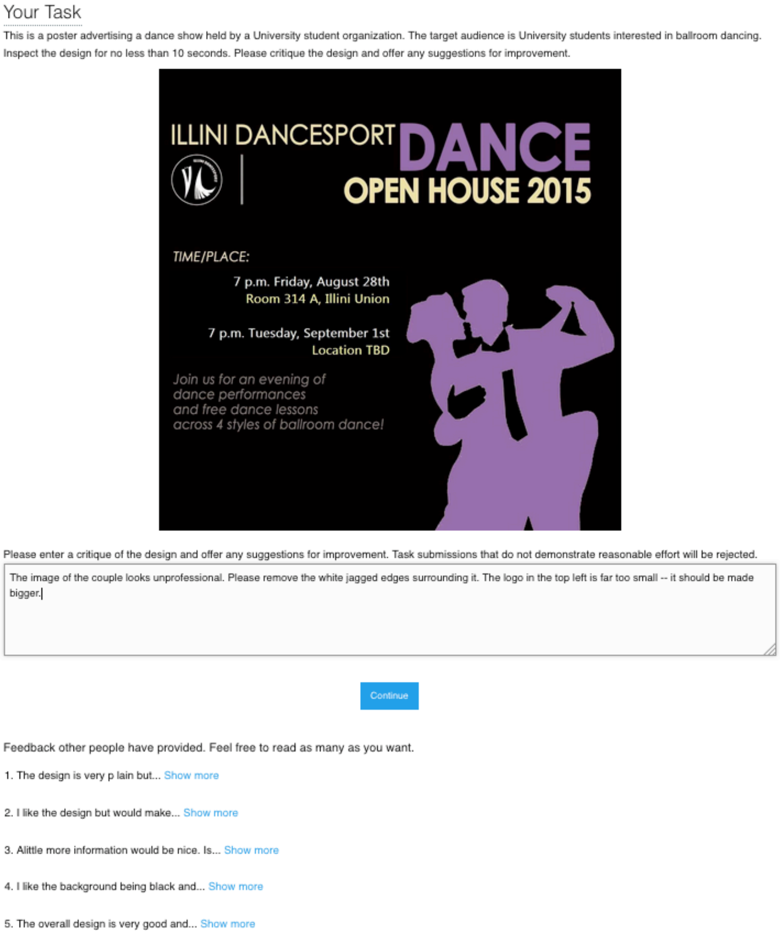


Figure 3.1. The chosen designs. Designs were selected to span a broad range of visual domains.

3.1.3 Feedback Interfaces

The feedback interface features a block of text introducing the task and includes a brief description of the design and its target audience. The design is then prominently displayed. Figure 3.2 introduces the non-spatial Interface. In this interface, a text area prompting the provider for feedback was below the design image. A submit button was placed next to the text area to complete the task. Past feedback was displayed underneath this form in the presence of history. Rather than pre-generating history, we mimicked real world systems by allowing the history to grow organically from feedback submitted by previous providers. The presentation of the history was based on how online platforms such as Reddit or Dribble function, where the provider has access to an evolving history [7, 8]. We adapted this format however to include a “Show more” interaction which allowed us to log which pieces of feedback were viewed.


ILLINI DANCESPORT
DANCE
OPEN HOUSE 2015
TIME/PLACE:
7 p.m. Friday, August 28th
Room 314 A, Illini Union
7 p.m. Tuesday, September 1st
Location TBD
Join us for an evening of dance performances and free dance lessons across 4 styles of ballroom dance!
Continue

Your Task

This is a poster advertising a dance show held by a University student organization. The target audience is University students interested in ballroom dancing. Inspect the design for no less than 10 seconds. Please critique the design and offer any suggestions for improvement.

Please enter a critique of the design and offer any suggestions for improvement. Task submissions that do not demonstrate reasonable effort will be rejected.

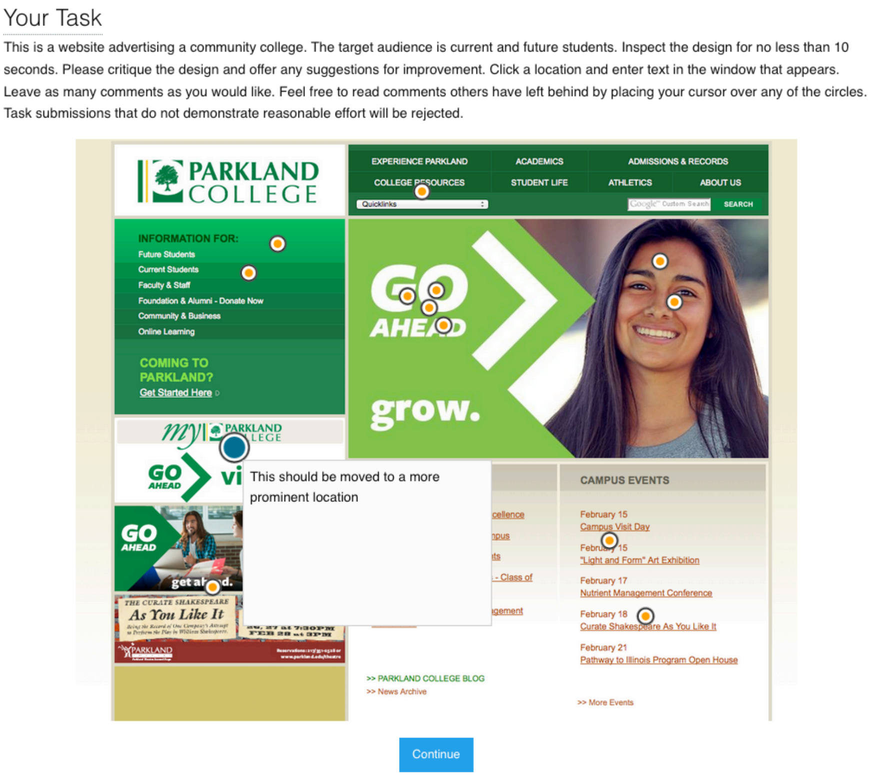
The image of the couple looks unprofessional. Please remove the white jagged edges surrounding it. The logo in the top left is far too small -- it should be made bigger.

Feedback other people have provided. Feel free to read as many as you want.

1. The design is very p lain but... Show more
2. I like the design but would make... Show more
3. Alittle more information would be nice. Is... Show more
4. I like the background being black and... Show more
5. The overall design is very good and... Show more

Figure 3.2. The interface for leaving feedback in the non-spatial condition. A feedback provider enters their feedback in a text area. In the history condition, feedbacks left by previous providers were visible. The participant may choose to view the full feedback by selecting “Show more.”

Figure 3.3 introduces the spatial Interface. In this interface, the feedback provider first selects a location on the design and is then prompted to enter feedback in the window that appears. The feedback is committed by pressing elsewhere on the image. To represent the feedback left at the location, a visual marker is overlaid on the design. As many pieces of feedback as desired could be entered. Providers could inspect the feedback they had left by hovering over the associated visual marker and could always edit their own feedback by clicking the marker. Instances of feedback left by the previous providers were shown in the presence of the History. The participant was allowed to hover over any visual markers to reveal the annotated feedback. The spatial condition was designed and implemented to reflect popular annotation feedback tools such as Adobe Acrobat and Red Pen [4, 17]. Once satisfied with the feedback, the provider submitted their work.



Your Task

This is a website advertising a community college. The target audience is current and future students. Inspect the design for no less than 10 seconds. Please critique the design and offer any suggestions for improvement. Click a location and enter text in the window that appears. Leave as many comments as you would like. Feel free to read comments others have left behind by placing your cursor over any of the circles. Task submissions that do not demonstrate reasonable effort will be rejected.

This should be moved to a more prominent location

Figure 3.3. The interface for leaving feedback in the spatial condition. A Provider can leave a comment by selecting a region on the design and entering text in a window. They could leave as many comments as desired and were allowed to look at the history by hovering existing markers.

3.1.4 Procedure

Upon accepting the task, the feedback provider was presented with a consent form. If accepted, the provider was randomly assigned to one of twelve experimental conditions. The conditions were implemented in JavaScript and the feedback provider did not have to leave the Mechanical Turk platform. In each condition, they read the task instructions, viewed the design, and entered feedback based on the interface condition assigned. After entering feedback in the interface provided, they submitted their work and completed a brief survey.

3.2 Measures

The study consisted of three sets of measures: content analysis, behavioral measures, and self-assessment.

3.2.1 Content Analysis

For content analysis, we calculated specificity, categorized the feedback content, and measured general metrics such as its length.

For each feedback response, a measure of specificity was calculated. Specificity was measured using the NLTK toolkit. The toolkit calculated specificity by determining how deep each word appears in the Wordnet structure. Words closer to the root are more general (e.g. “dog”) while deeper words are more specific (e.g. “Labrador”). Stop words and punctuation were ignored. The specificity metric was normalized to range from 0.0 to 1.0. In the past, other researchers have used this technique [20].

We categorized the feedback content by classifying the individual idea units that compose responses. Each feedback response was partitioned into individual idea units. An idea unit represents a coherent unit of thought. The idea units were then coded based upon a taxonomy of critique discourse [21]. For example, the taxonomy included categories for judgement (*"I like that sketch but not that design. I don't like this up here because it looks paperish—you know, not ceramic."*) and interpretation (*"There's a whole mysterious quality. There's a shadow and a mystery, and you wonder, what's going on in there?"*).

Two coders with experience in HCI used the taxonomy to categorize each idea unit. In total, 1206 idea units were categorized. Cohen's Kappa, a measure of reliability between multiple raters, was 0.81 on 80 training samples (5% of the dataset). Coders were paid $25 for their effort.

Additionally, we measured feedback length by cumulative character count of all feedback from a single provider.

3.2.2 Behavioral Measures

For behavioral measures, we calculated the similarity between generated feedback and history feedback and computed general behavioral metrics.

A provider's interactions with prior feedback were logged. For the spatial condition, we logged each time the provider revealed a previous feedback by hovering over a visual marker. Likewise, in the non-spatial condition, we logged each time the provider selected a "Show more" link.

We used a distance metric to calculate whether each comment that a provider left was more similar to history feedback that was reviewed. To compile the reviewed feedback, we aggregated the set of history that they had viewed for at least one second. The history they had not reviewed was also aggregated. To measure similarity, the distance between the recent comment and feedback that was and was not reviewed was calculated using the Python pattern.en toolkit.

We also measured behavioral metrics such as task completion time, the number of prior feedback responses revealed, and the count of feedbacks provided. These measures help us understand how different interface conditions affected the behavior of feedback the provider.

3.2.3 Self-assessment

Following the feedback task, the provider completed a self-assessment survey. They rated their design expertise, perceived effort, and the perceived usefulness of given feedback on a five point Likert-scale, with a score of 5 as the most favorable. The survey also included two questions for demographics (age and gender).

CHAPTER 4

RESULTS

In total, 30 responses were collected per experimental condition for a total of 360 responses. We reviewed all the submissions and excluded any that were irrelevant or incomprehensible. Three submissions were excluded, leaving us with 357 feedback responses of reasonable quality.

4.1 Content Analysis

To analyze the content, we calculated feedback specificity, categorized the feedback content, and measured general metrics such as its length.

4.1.1 Non-spatial condition produced longer feedback

An ANOVA revealed that Interface had a main effect on feedback length (F(3,357)=7.86; *p*=0.0053). Character count per condition can be seen in Figure 4.1.

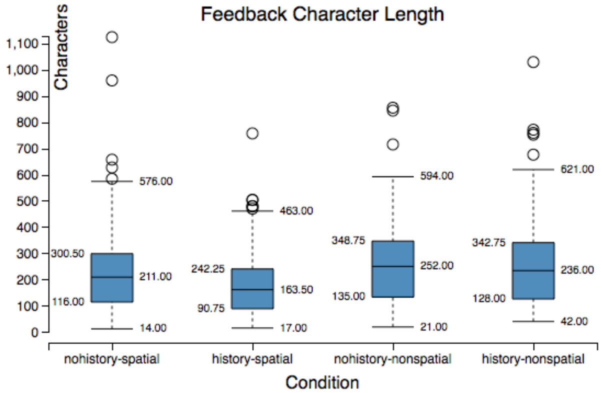


Figure 4.1. The effect of experimental condition on length of feedback content is shown in this chart. Analysis shows providers left longer feedback in the non-spatial condition. No other effects were found.

Pairwise comparison using Tukey's HSD showed that the length of the feedback in the non-spatial condition (μ=269.7 characters) was longer than the feedback from the spatial condition (μ=217.4; p=0.0051). No other effects were discovered.

The non-spatial condition may have led to longer feedback due to the need for use of deixis, i.e. words or phrases such as "here" or "there" that require further contextual information to be understood but eliminate the need for explicit description of the visual elements referenced by feedback.

4.1.2 Conditions produce different categories of feedback

After categorizing the idea units from generated feedback, we performed z-tests for population proportions to look for patterns of interest.

Table 4.1 shows the breakdown of idea unit categories per condition. We found that the spatial Interface generated more *investigations* (4.1%) than the non-spatial Interface (1.1%; z=3.23; p=0.001).

	Condition				
	Non-spatial		**Spatial**		
Category	**No History**	**History**	**No-History**	**History**	**Total**
Judgement	47.2% (151)	53.8% (164)	44.9% (151)	44.1% (154)	620
Recommendation	39.4% (126)	32.1% (98)	29.5% (99)	36.1% (126)	449
Investigation	1.2% (4)	1.0% (3)	4.5% (15)	3.2 % (11)	33
Interpretation	2.8% (9)	0.7% (2)	2.1% (7)	1.7% (6)	24
Brainstorming	4.7% (15)	6.9% (21)	10.4% (35)	5.2% (18)	89
Process	0.3% (1)	0.0% (0)	0.0% (0)	0.0% (0)	1
Comparison	0.9% (3)	1.0% (3)	1.5% (4)	2.0% (7)	17
Identity Invoking	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	0
Association	1.9% (6)	1.6% (5)	0.0% (0)	1.4% (5)	16
Total Idea Units	**315**	**296**	**311**	**327**	**1249**

Table 4.1. Frequencies of the categories of idea units by Interface and History.

An *investigation* is when the feedback provider asked questions about a specific piece of the design. Referencing specific pieces of the design may have been encouraged by the spatial interface requiring the provider to visually scan elements to select a location on the design before entering feedback.

No other significant results in category frequency were discovered.

4.1.3 Non-spatial feedback had more stop words

An ANOVA did not detect a main effect of Interface or History on feedback specificity. In the spatial condition, mean specificity was 0.34 ($\sigma = 0.17$), while the non-spatial condition had a mean specificity of 0.37 ($\sigma = 0.14$).

An ANOVA uncovered a main effect of Interface on stop word count (F(3,357)=6.93; p=0.0089). Figure 4.2 summarizes stop word count. Tukey's HSD showed that stop word count in the non-spatial condition (μ=27.31) was greater than the spatial condition (μ=21.98; p=0.0084).

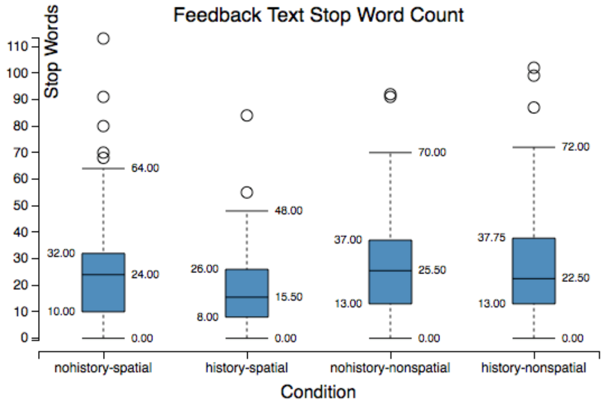


Figure 4.2. This chart shows how the experimental condition affected stop word count of the feedback content. Analysis shows providers included more stop words in their feedback in the non-spatial condition.

Higher stop word count in the non-spatial condition may be attributed to the interface's lack of context relative to the spatial interface. The additional context provided by the spatial condition reduced the need for language necessary to convey the same information as the non-spatial interface. In the non-spatial condition, stop words were used to reference specific elements of the design: *"The logo must come at top before title and it must be large. The sentence written at the bottom should be brightened… There should be a name and contact details of a person to contact."* Providers neglected these words in the spatial condition: *"Unappealing shade of purple. Perhaps more distinctness between the two silhouettes – looks kind of blobby right now. Maybe use bullet points."*

4.2 Behavioral Measures

To analyze provider behavior, we calculated the similarity between generated feedback and reviewed history feedback and computed general behavior metrics such as task completion time, number of prior responses reviewed, and count of feedbacks provided.

4.2.1 Providers inspect more feedback in the spatial Interface

When History was present in the spatial Interface, we found 55% of providers (99 providers) inspected history feedback. In the non-spatial Interface, we found only 19% of providers (33 providers) inspected history feedback. The number of instances of feedback inspected by providers is visualized in Figure 4.3. An ANOVA revealed a main effect of Interface on instances of feedback inspected (F(3, 180)=60.57; p=0.0001). Tukey's HSD showed that spatial Interface providers inspected more feedback instances (μ=7.29) than the non-spatial condition (μ=1.14, p=0.0001).

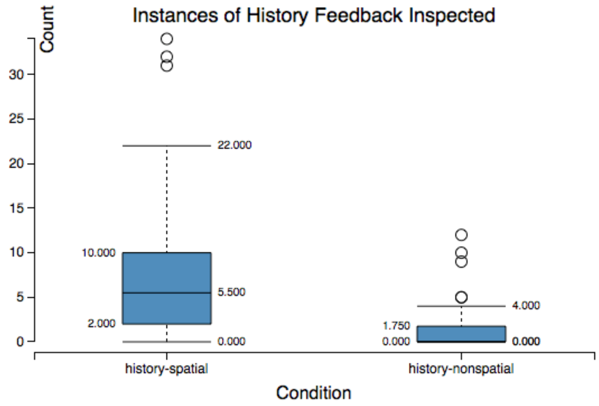


Figure 4.3. This chart shows the count of instances of History condition feedbacks inspected by Interface. Analysis shows providers inspected more feedback instances under the visual condition.

One explanation for this effect is the cost of access of history feedback in the non-spatial Interface relative to the spatial interface. Providers in the spatial Interface didn’t have to scroll and didn’t have to click a “Show more” link to unveil history feedback.

4.2.2 Generated feedback was more similar to viewed history

We only considered instances of generated feedback where the provider had reviewed some history. This left us with 200 instances of feedback in the spatial condition and 42 instances of feedback in the non-spatial condition. Figure 4.4 displays feedback similarity scores.

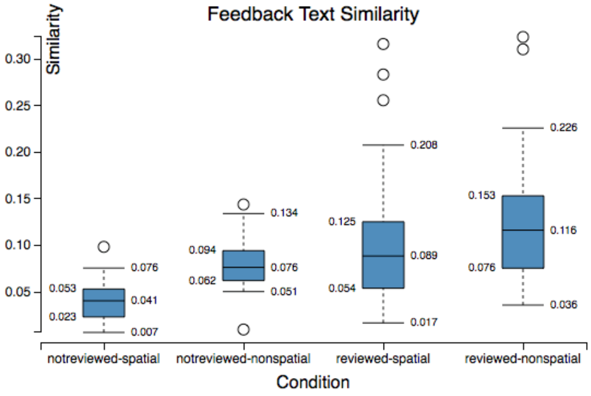


Figure 4.4. Similarity scores of generated feedbacks compared to viewed and unviewed history by condition. Generated feedback was more similar to viewed history.

An ANOVA showed that when a provider generated feedback, the feedback was more similar to the history that the provider reviewed (μ=0.11) than it was to the history that the provider did not review (μ=0.044; F(3,232)=26.59; p=0.0001). Tukey's HSD showed this difference significant (p=0.0001).

This suggests that presence of a History exposes feedback providers to a fixation effect. This effect is analogous to how pictorial representations of examples introduce a fixation effect when solving design problems [15].

4.2.3 Non-spatial feedback was more similar to viewed history

An additional main effect revealed by ANOVA was the influence of Interface on similarity to viewed history (F(3,230)=12.88; p=0.0004). Tukey's HSD showed that similarity to reviewed feedback in the non-spatial condition (μ=0.11) was higher than that of the spatial Interface (μ=0.069; p=0.0039). This effect is visible in Figure 4.4.

The prominence of the fixation effect in the non-spatial condition may have been due to the more permanent nature of revealed history. Revealed history in the non-spatial interface remained visible until the provider explicitly chose to hide it. Meanwhile, in the spatial interface, providers had to continuously hover over a marker to reveal the feedback content.

Analysis of data did not show effects of conditions on task completion time. Providers completed the task in 221.3 seconds on average (σ=178.27 seconds).

4.3 Self-Assessment

Feedback providers self-reported their design expertise, perceived effort, and the perceived usefulness of given feedback.

4.3.1 Design influenced perceived usefulness of the feedback

Table 4.2 shows the breakdown of effort ratings across conditions. ANOVA did not detect differences between these conditions.

An ANOVA detected a main effect of Design on self-assessed feedback usefulness rating ($F(3,357)=5.0$; $p=0.046$). Perceived usefulness of the feedback generated in Design B ($\mu=4.1$; $\sigma=0.86$) and Design C ($\mu=4.0$; $\sigma=0.89$) was higher on average than that of Design A ($\mu=3.8$; $\sigma=0.93$).

An explanation for this effect is the fact that Design A had more opportunity for improvement since it was designed by a novice, whereas Designs B and C were professional web pages.

Effort Self-assessment		
	Non-spatial	**Spatial**
History Absent	$\mu=3.3$; $\sigma=1.1$	$\mu=3.1$; $\sigma=1.2$
History Present	$\mu=3.3$; $\sigma=1.2$	$\mu=3.1$; $\sigma=1.0$

Table 4.2. Provider perceived effort self-assessment by condition. Conditions had no significant effect of perceived effort.

CHAPTER 5

DISCUSSION AND FUTURE WORK

The goal of our work was to study the influence of Interface and History on generated feedback. Providers in the spatial interface reviewed more history feedback. We found that the presence of a History introduced a fixation effect. This effect caused feedback providers to generate feedback that is more similar to the history they reviewed. Convergent responses are encouraged when a designer chooses to use a history enabled tool to generate feedback. This effect was more prominent in the non-spatial interface.

We found that providers using the non-spatial interface produced feedback that was 24% longer and had more stop words. Conditions produced different categories of feedback. The spatial Interface generated more *investigation* feedback. This category of feedback may be particularly useful at the early stages of design. Our results did not detect differences in the frequency of other categories of feedback. The conditions had no impact of the specificity of generated feedback.

While Interface and History did not influence self-assessed perceived feedback usefulness, providers critiquing the novice design tended to perceive their feedback as less useful. Results found no interaction between conditions and self-assessed provider effort.

Our results did not detect differences in task completion time between conditions.

We conclude that the only factors relevant to the decision over feedback interface are desired feedback length and category. Maybe the more important choice designers must make is not the class of tool, but whether the history is included. The reason these results are interesting is designers can have more confidence that their choice of interface for providers will have little consequence for the feedback they collect.

Designers seek different kinds of feedback at different stages in the design process. For example, low-fidelity paper prototypes encourage early exploration of more design alternatives [13, 22]. Different stages of design and their interaction with choice of design tool were not considered in our study. It was not feasible to include an additional factor of design stage due to the number of factors we studied.

The conditions in our study represented two classes of feedback interface with features that were exclusive of each other. However, there are feedback interfaces that do not fit neatly into either of these conditions. The Voyant and CrowdCrit systems, in addition to offering a free-form response and a text box, also allowed the provider to annotate a region of the design to associate with the comment [3, 10]. Future work is necessary to understand how these hybrid interfaces compare to the two conditions that we studied.

We also did not consider different levels of expertise of the feedback provider. For instance, an expert may find it less necessary to access the history of feedback when generating their own insights [10]. Experts also tend to both generate more ideas and to fixate more often [15]. On the other hand, novices may value access to history feedback for inspiration [20]. Exploring how expertise interacts with the choice of feedback collection interface will require future work.

5.1 Limitations

We studied the influence of interface and history on features of the feedback generated. However, the feedback objective quality was not measured. Evaluating the quality of the feedback by recruited independent experts could address this limitation in future work.

A second limitation is that we recruited feedback providers from the Mechanical Turk platform. Workers recruited from this platform were primarily incentivized by financial gain.

Crowds driven by different incentives, such as classroom peers, people recruited from social networks, or in the context of online communities such as Reddit could be studied in future work to test the generalizability of these findings.

CHAPTER 6

CONCLUSION

Designers are increasingly turning to a variety of online tools for collecting feedback on in-progress work. The main contribution of this work is reporting how interface and history influence the generated feedback. First, we showed that access to history across both classes of interfaces studied introduced a fixation effect. Fixation can therefore lead to more convergent feedback or feedback that has less diversity. Diversity is one of the reasons people want to collect feedback from a large audience online. Second, we found that providers in the non-spatial interface generated 24% longer feedback. But while generated feedback in this condition is longer, it is not necessarily more useful. Third, we discovered that spatial interfaces generate more *investigation* feedback. While this result is significant, there was no effect of conditions on the frequency of the other nine categories of feedback. We hope this work contributes to more extensive interfaces for feedback exchange, helps feedback providers better communicate the evaluation of the design, and aids designers in making effective decisions when selecting feedback tools.

REFERENCES

1. Elkins, J., *Art Critiques: A Guide.* New Academia Publishing, 2012.

2. Xu, A., et al., *A Classroom Study of Using Crowd Feedback in the Iterative Design Process*, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*. 2015, ACM: Vancouver, BC, Canada. p. 1637-1648.

3. Xu, A., S.-W. Huang, and B. Bailey, *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 &#38; social computing*. 2014, ACM: Baltimore, Maryland, USA. p. 1433-1444.

4. Willett, W., J. Heer, and M. Agrawala, *Strategies for crowdsourcing social data analysis*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, ACM: Austin, Texas, USA. p. 227-236.

5. *Adobe Acrobat.* Available from: https://get.adobe.com/reader/.

6. *Red Pen.* Available from: https://redpen.io/.

7. *Reddit*. Available from: https://www.reddit.com/r/design_critiques.

8. *Dribble.* Available from: https://dribbble.com/.

9. Hicks, C.M., et al., *Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment*, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, ACM: Santa Clara, California, USA. p. 458-469.

10. Luther, K., et al., *Structuring, Aggregating, and Evaluating Crowdsourced Design Critique*, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*. 2015, ACM: Vancouver, BC, Canada. p. 473-485.

11. Greenberg, M.D., M.W. Easterday, and E.M. Gerber, *Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers*, in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 2015, ACM: Glasgow, United Kingdom. p. 235-244.

12. Yu, L. and J.V. Nickerson, *Cooks or cobblers?: crowd creativity through combination*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011, ACM: Vancouver, BC, Canada. p. 1393-1402.

13. Tohidi, M., et al., *Getting the right design and the design right*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2006, ACM: Montr&#233;al, Qu&#233;bec, Canada. p. 1243-1252.

14. Siangliulue, P., et al., *Providing Timely Examples Improves the Quantity and Quality of Generated Ideas*, in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 2015, ACM: Glasgow, United Kingdom. p. 83-92.

15. Gero, J.S., *Fixation and Commitment While Designing and its Measurement.* The Journal of Creative Behavior, 2011. **45**(2): p. 108-115.

16. Viswanathan, V.L., Julie, *UNDERSTANDING FIXATION: A STUDY ON THE ROLE OF EXPERTISE.* Proceedings of the 18th International Conference on Engineering Design (ICED 11), 2011. **7**: p. 309-319.

17. Hill, W.C. and J.D. Hollan, *Deixis and the future of visualization excellence*, in *Proceedings of the 2nd conference on Visualization '91*. 1991, IEEE Computer Society Press: San Diego, California. p. 314-320.

18. Yoon, D., et al., *RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion*, in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016, ACM: San Francisco, California, USA. p. 195-205.

19. Robb, D.A., et al., *Moodsource: Enabling Perceptual and Emotional Feedback from Crowds*, in *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work &#38; Social Computing*. 2015, ACM: Vancouver, BC, Canada. p. 21-24.

20. Yuan, A., et al., *Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques*, in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016, ACM: San Francisco, California, USA. p. 1005-1017.

21. Dannels, D.P., and Martin, K. N., *Critiquing critiques a genre analysis of feedback across novice to expert design studios.* Jo. Bus. & Tech. Comm. , 2008. **22**(2).

22. Rettig, M., *Prototyping for tiny fingers.* Commun. ACM, 1994. **37**(4): p. 21-27.