

COMPARING TEXT AND VISUAL ANNOTATION TOOLS FOR DESIGN FEEDBACK

BY

ANTOINE DEJONG

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, 2017

Urbana, Illinois

Adviser:

Professor Brian P. Bailey

ABSTRACT

Designers rely on critiques to develop their skills and iterate toward effective designs. Designers are increasingly turning to online tools and communities to collect affordable, scalable feedback. There are many tools available but little empirical evidence to guide the decision to select one over another. We conducted a study on Amazon Mechanical Turk (N=360) to contrast two popular classes of feedback collection interfaces: text and spatial. The text interface has one text box for providers to submit their feedback, whereas the spatial interface allows the providers to annotate the design. We also manipulated the presence and content of history of feedback visible to providers. The three history conditions were aesthetic, goal-related, and ‘no-history’ if they did not have to option to view feedback. We found that the presence of sample feedback induces a fixation effect in design reviews, causing feedback to be more similar to the feedback that the providers reviewed. We found that both the Interface and History conditions have statistically significant effects on the content of feedback provided. However, neither resulted in greater perceived quality of feedback. Our study found that feedback in the text condition was 27% longer than in the spatial condition. We also found that providers who reviewed goal-oriented sample feedback rated the task as being easier than providers who did not receive sample feedback. These results indicate that the more important decision for designers is not which class of interface to use but whether to include history or sample feedback.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
CHAPTER 2: RELATED WORK.....	6
CHAPTER 3: METHODOLOGY	14
CHAPTER 4: RESULTS	25
CHAPTER 5: DISCUSSION.....	32
REFERENCES	40

CHAPTER 1

INTRODUCTION

Critiques are an integral part of the design process. Students require feedback to develop their skills. Professional designers need feedback to iterate towards solutions that satisfy user needs [1]. However, the process of securing feedback from experts can be made difficult by limitations on time, money, and access to trained experts. Designers are increasingly turning to online feedback platforms to circumvent these limitations. Online platforms can aid in gathering feedback in a manner which is scalable, fast, and affordable [2-4]. There is a variety of online tools for feedback collection. However, it is unclear which tool a designer should select and under what circumstances.

One way in which feedback collection tools differ is in their user interfaces. Two of the most common classes are text-centric and spatial interfaces. For text interfaces, the feedback provider enters their comments into a single text area. In this case, the feedback is text focused and open ended. This requires providers to reference the whole design, which may cause them to consider general design principles. Some examples of this type of interface are Dribbble and Reddit [5, 6]. Spatial interfaces, such as Redpen.io, allow providers to markup the document by placing feedback directly onto the design. This involves the provider visually scanning the design for areas to place comments. This may cause providers to become more focused on particular aesthetic features, rather than conceptual issues [2, 3]. Examples include Adobe Acrobat and Redpen.io [7, 8]. Others, such as Voyant and CrowdCrit, involve a mix of the two, allowing providers to both annotate the design and to write additional comments [2, 9].

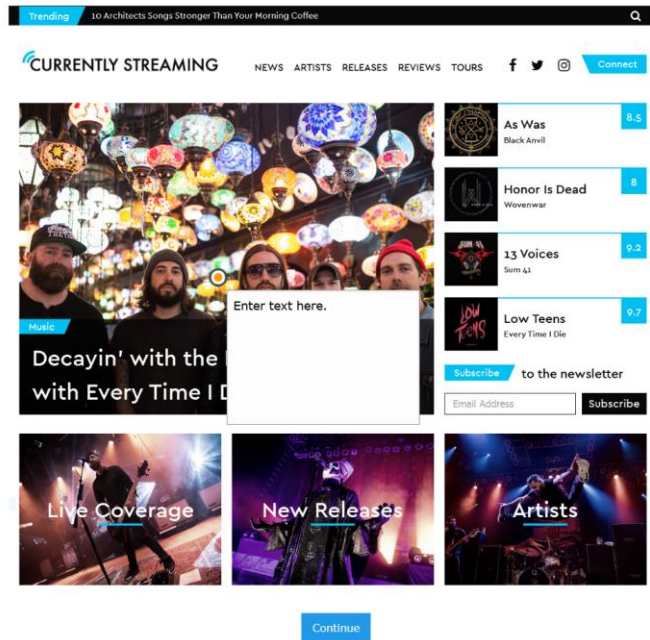


Figure 1.1. An example of a spatial interface. The feedback provider chooses points on the design to place annotations.

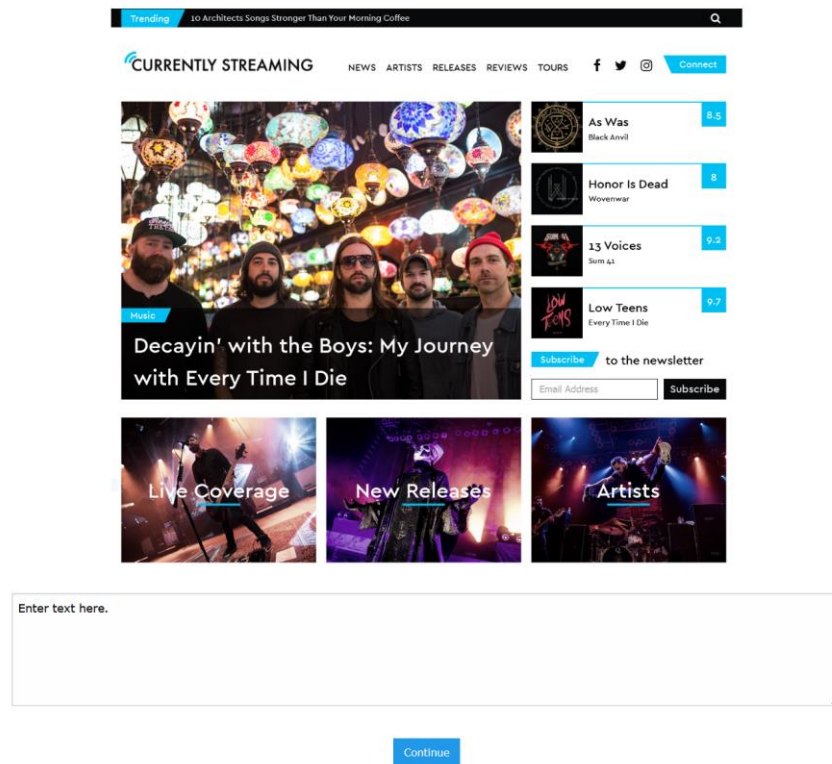


Figure 1.2. A text interface. All feedback is entered through the text box.

In this study, we compare spatial and text focused interfaces. Changes in the feedback interface are known to have significant impacts on the feedback generated. For example, scaffolding the task of providing feedback improves the quality of feedback [4, 10]. However, there are not yet widely established best practices for gathering design feedback online. Investigating the differences between common types of interface will help to inform designers which platforms will best suit their needs.

Fixation has long been recognized in the study of design. When tasked with creating visual or product designs, the presence of examples causes designers to become fixated, generating designs which share features with the examples [11]. This effect has been observed even for features which are detrimental to the design [12]. The study of fixation has focused on the creative process. Although fixation is well established, it is unknown whether fixation could also occur in more analytic tasks such as the process of generating feedback.

For some websites used for feedback collection (e.g., Reddit) the history of feedback left by previous providers is visible to later providers. This may impact the later reviewers. It is unknown whether this would have an effect which improve creativity [13-15] or limit divergent feedback by introducing a fixation effect [16, 17].

To examine the influence of showing feedback history, we obtained sample feedback from a pilot study and sorted the feedback according to whether the example referenced meeting the designer's goals and objectives (goal-oriented feedback) or not (aesthetic feedback). It has been found that dividing the feedback process into smaller, focused tasks leads reviewers to generate more goal-oriented feedback [18]. Similarly, showing goal-oriented examples might encourage reviewers to give more conceptual feedback. We expected that goal-oriented feedback would help

to frame the feedback generation in the context of the designer's motivations. This could inspire providers to reflect more deeply on the design and thus generate more useful feedback.

The study involved a total of six conditions. We included two feedback collection interfaces, text and spatial. For each interface, we manipulated the presence and content of History in three conditions: goal, aesthetic, or no history.

We recruited feedback providers from Amazon Mechanical Turk (N=360). For each condition, providers gave feedback on one of two designs. Providers were also presented a brief explanation of the goals of the design. Providers were assigned to an interface to enter their feedback. In the goal and aesthetic History conditions, participants had the option to review sample feedback.

We measured the length of feedback provided. We assessed the similarity of feedback to viewed history as the cosine similarity TF-IDF vectors, a widely accepted measure of textual similarity [19]. Provider feedback was divided into individual idea units and coded according to category as in Cho et al [20], as well as by topic. The quality of each idea unit was also rated from 1-5. Participants completed a brief self-assessment and demographics survey after submitting their feedback.

The study found that feedback is 27% longer for the text interface compared to the spatial interface. The history condition had an impact on the category of the idea unit and the topic of discussion, but there was no significant difference in perceived quality. We also asked providers to rate the amount of effort they felt the task required on a scale of 1-5. We found that reviewers with access to goal-oriented history rated the task as requiring less effort (3.1) than those in the

no-history condition (3.5, $p=0.02$). This was not the case with providers with access to aesthetic samples.

Our findings indicate that the presence of a visible history induces a fixation effect. Providers left feedback which was more similar to example history feedback they had reviewed ($\mu=0.23$) than it was to history which providers had not reviewed ($\mu=0.20$, $p<0.0001$). Similarity was measured as the cosine similarity of TF-IDF vectors in a range from 0-1, a widely accepted measure of text similarity [19]. This causes feedback to be less diverse, potentially limiting its utility. Designers looking for independent opinions on their work should choose tools or platforms which prevent reviewers from seeing each other's contributions before submitting their own. Designers who want to gather diverse, independent feedback should not give providers access to previous feedback.

CHAPTER 2

RELATED WORK

Feedback collection tools have been used in the categories of education, design, and for gathering feedback from the crowd.

2.1 Education Feedback Collection Tools

Annotation has been studied extensively in educational contexts. A focus of the study of annotation in education is social annotation, using online tools to annotate class materials [21]. The purpose of these is enrich discussion between students and to give the instructor an impression of the class' understanding. Two recent examples are NB and Mudslide. NB is a social annotation tool which allows students to review and comment on class readings [22]. Students can interact in real time as they read the material in a threaded discussion.

Mudslide [23] is a visual tool which allows students to point out areas of confusion in lectures by marking up a copy of the lecture slides. Each student may paint sections on slides which correspond to areas of personal confusion. When viewed in the aggregate by the instructor, each student's markers are treated with an opacity layer so that darker, more visible marks indicate areas of confusion for several students. In the interest of scalability, students do not submit text but markup the lecturer's slides. In this way, students may identify problem areas generally, but they are unable to give clarifying information or directly convey information about the source of the confusion. Markers were also color coded red if the student found the lecture was confusing overall or gray if they did not. This forms a heat map for teachers to evaluate the class's understanding and to address points of confusion.

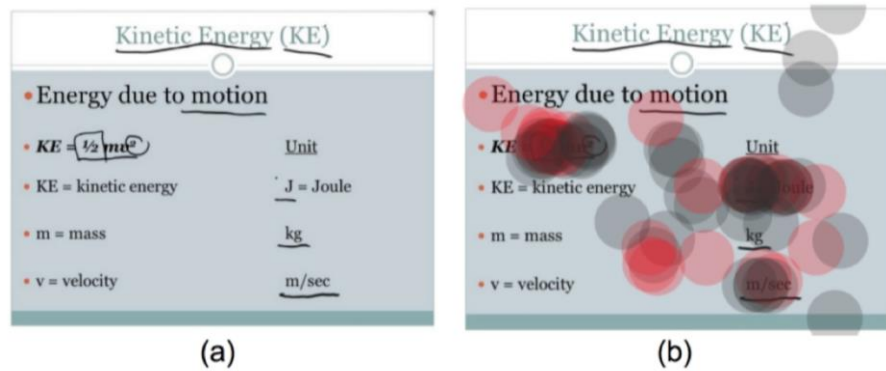


Figure 2.1. Example of Mudslide. (a) shows the final lecture slide and (b) shows the slide with "muddy" annotations for points of confusion from the students. Borrowed from Figure 4 of "Mudslide"

RichReview++ [24] is part of the class of spatial tools. It is a multimodal document annotation that allows users to not only place comments, but also to leave voice recordings, transcriptions, and gestures. Participants using RichReview++ found the features useful and found it aided in navigating documents. Students also stated a preference for RichReview++ over traditional feedback methods, such as face-to-face interactions.

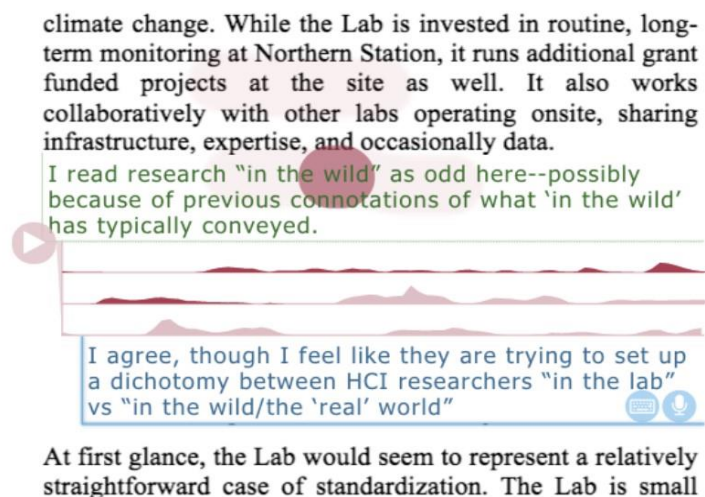


Figure 2.2. The RichReview++ interface. RichReview++ allows providers to leave feedback in the modalities of text, voice, and gestures. From Figure 1 of "RichReview++"

2.2 Design Feedback Collection

There are multiple methods for gathering design feedback. The interface for directly gathering feedback may be spatial, textual, or a mixture of the two. In spatial interfaces, feedback providers are asked to tie their feedback to a specific point or selection of the design. This requires the provider to visually scan the design for areas to place their feedback. It may encourage providers to focus on specific visual elements, rather than considering the design holistically. Widely used applications including Microsoft Word and Adobe Acrobat have implemented spatial interfaces for their annotation systems. Redpen.io is an example of the spatial interface as a collection tool for design feedback. In Redpen.io, providers place markers onto the design and enter text for each of these markers. See figures 2.3 and 2.4.

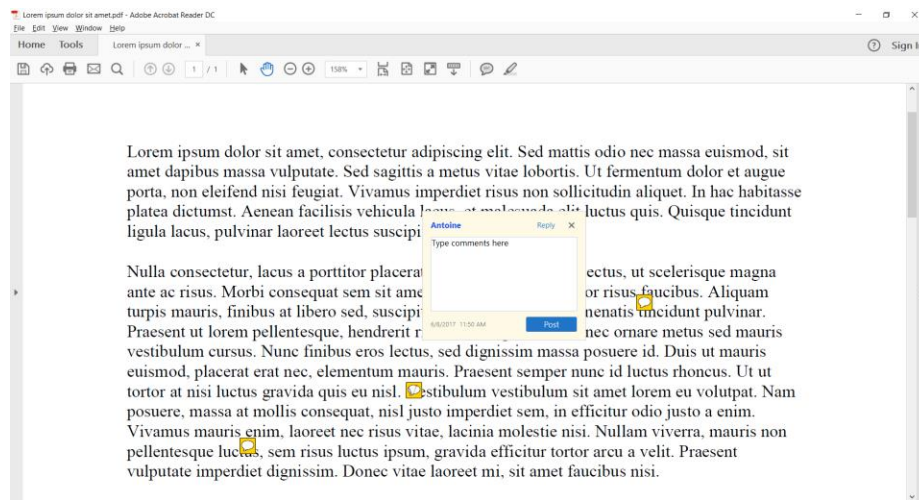


Figure 2.3. Adobe Acrobat comments, an example of a spatial feedback interface. Feedback providers must select a location to place the comments before they may enter feedback. This requires them to visually scan the design. This may cause providers to focus on aesthetic details.

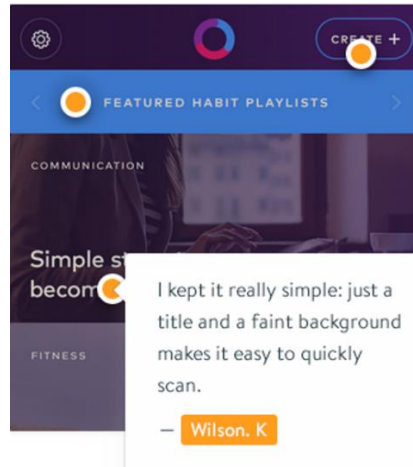


Figure 2.4. The Redpen.io feedback interface. Designers and feedback providers place a marker before leaving text. This requires visually scanning the design, which may influence the nature of feedback. From redpen.io.

The other form of interface for entering feedback is through a text interface. In this style of interface, a text box is presented alongside the design and all feedback is entered into this text box. Subforums of the website Reddit such as r/design_critiques are home to online communities which share designs. Reddit displays the design or a link to it and providers may enter their feedback in a text box just beneath it. Providers may also review prior feedback that has been left by others. Such websites may generate high quality feedback but require social capital [25] and generally receive few responses [3]. Dribbble and Behance are websites which are dedicated feedback collection platform. Like Reddit, they collect open-response feedback through a text box presented below a design.

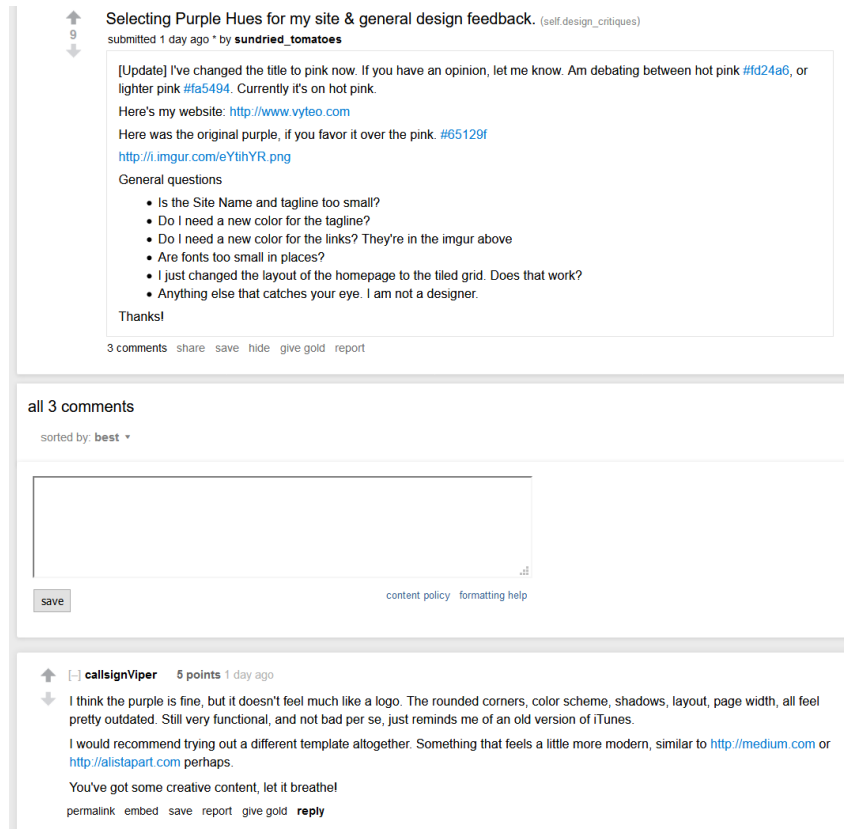


Figure 2.5. The Reddit feedback interface. A text box is displayed prominently below a link to the design. Providers may also review prior feedback.

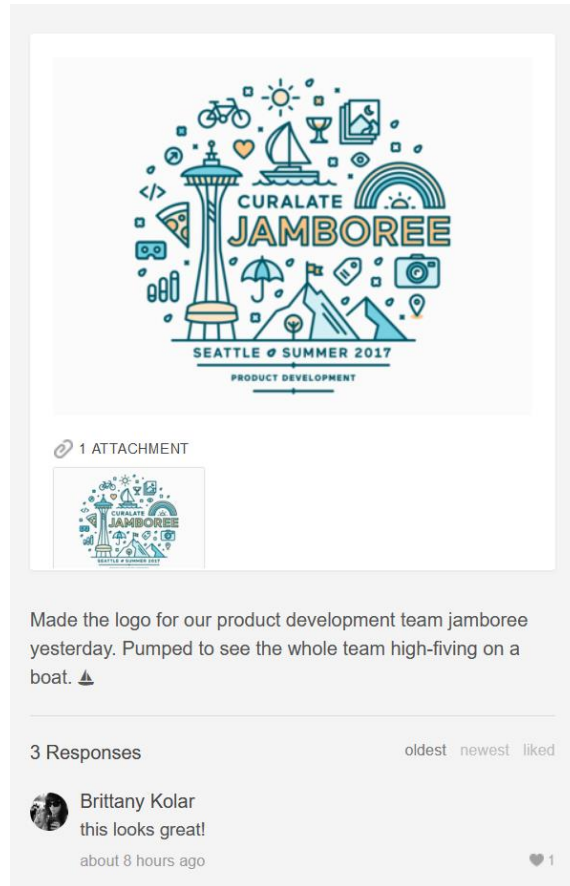


Figure 2.6 The feedback interface for Dribbble. Dribbble is an example of a text interface.

2.3 Crowd Feedback Platforms

Crowd feedback platforms differ somewhat from educational tools and because there is generally less interaction between reviewers. Rather, the emphasis is on collecting many critiques and presenting them to the designer in a scalable manner. Crowd feedback has been shown to improve designs [3] and to increase designer confidence [10]. Hicks et al [18] found that dividing tasks into narrower sub-tasks caused providers to give more goal-oriented feedback as opposed to purely aesthetic or other feedback.

Voyant is a design feedback tool in which designers select feedback categories of interest. Voyant would then automatically generated tasks and posted them to an online marketplace [2].

Task results were then aggregated and presented to the designer. The feedback was considered useful by designers, particularly the feedback providers' initial impressions and their responses to the designers' stated goals. Designers also found Voyant's organization and representation of feedback to be useful.

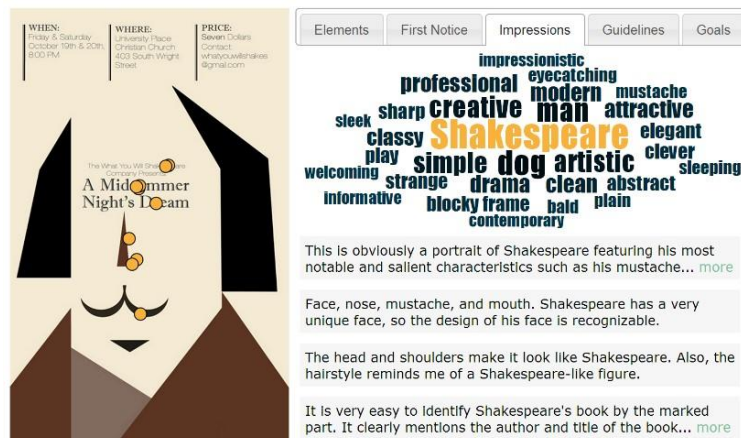


Figure 2.7. The interface of the Voyant crowd feedback platform. The interface is showing a word cloud of impressions from the crowd. Borrowed from Figure 1 of "Voyant"

CrowdCrit [9] helps generate feedback by scaffolding feedback providers. CrowdCrit has seventy prefabricated statements that providers use to construct their feedback. Feedback is then aggregated and displayed to the designer, who may explore the feedback. Aggregated crowd feedback was found to approach that of experts. Designers involved in the study were also favorably inclined towards the crowd feedback they received and felt that it had improved their designs, although there was not an observed improvement in designs after receiving crowd feedback compared to generic feedback.



Figure 2.8. The feedback collection interface for CrowdCrit. The platform scaffolds feedback by allowing providers to select from 70 prewritten comments. Comments are tied to an annotation on part of the website and providers may write additional clarifying comments. Borrowed from figure 3 of "CrowdCrit"

A platform called Critiki presents a simple solution for collecting reviews on crowdfunding project pages. Critiki also uses scaffolding to improve feedback quality [4], by showing high quality example points alongside worker prompts. Critiki was validated in two studies, with the first demonstrating that the crowd performed similarly to crowdfunding experts. In the second, crowdfunding designers reported that the platform was usable and that the feedback would be particularly useful to them in earlier stages of development. Critiki [4] also suggests that finding consensus in crowd feedback may be particularly useful to designers in the future.

CHAPTER 3

METHODOLOGY

This study compared how two classes of feedback Interface (text and spatial) and three classes of feedback History (goal, aesthetic, and no history) affect feedback generated in response to visual designs. We address the following questions: what is the impact of user interface on feedback? Does history induce a fixation effect? Is goal-oriented History more effective for guiding reviewers than aesthetic feedback? These questions are intended to aid designers in selecting feedback collection interfaces and knowledge of how that choice may affect the feedback received online.

3.1 Experimental Design

To answer these inquiries, we conducted a full-factorial, between-subjects experiment. The factors were Design category (poster vs. website) x Interface (text vs. spatial) x History (aesthetic vs. goal-oriented vs. no-history), resulting in a 2x2x3 design.

3.1.1 Participants

Feedback providers (N=360) were recruited from Amazon Mechanical Turk. Providers were required to have already completed at least 50 tasks and an overall task approval rate greater than 95%. 186 females and 173 males aged eighteen and over participated in the study. Providers were required to be located within the US. Based upon a pilot study, payment was set at \$0.70 per task to reflect the current local minimum wage.

3.1.2 Designs

We selected two designs from different visual domains to be both familiar to a general audience and to motivate suggestions for improvements. The poster design advertises a walking architecture tour and was found through the artist's collection on Behance. The website was a prototype for a site about music news and articles, and was posted to the “r/design_critiques” subreddit. Explicit permission was granted by the creator of the poster design. The other was used in accordance with the fair use principles of US copyright laws. The designs are shown in Figure 3.1.

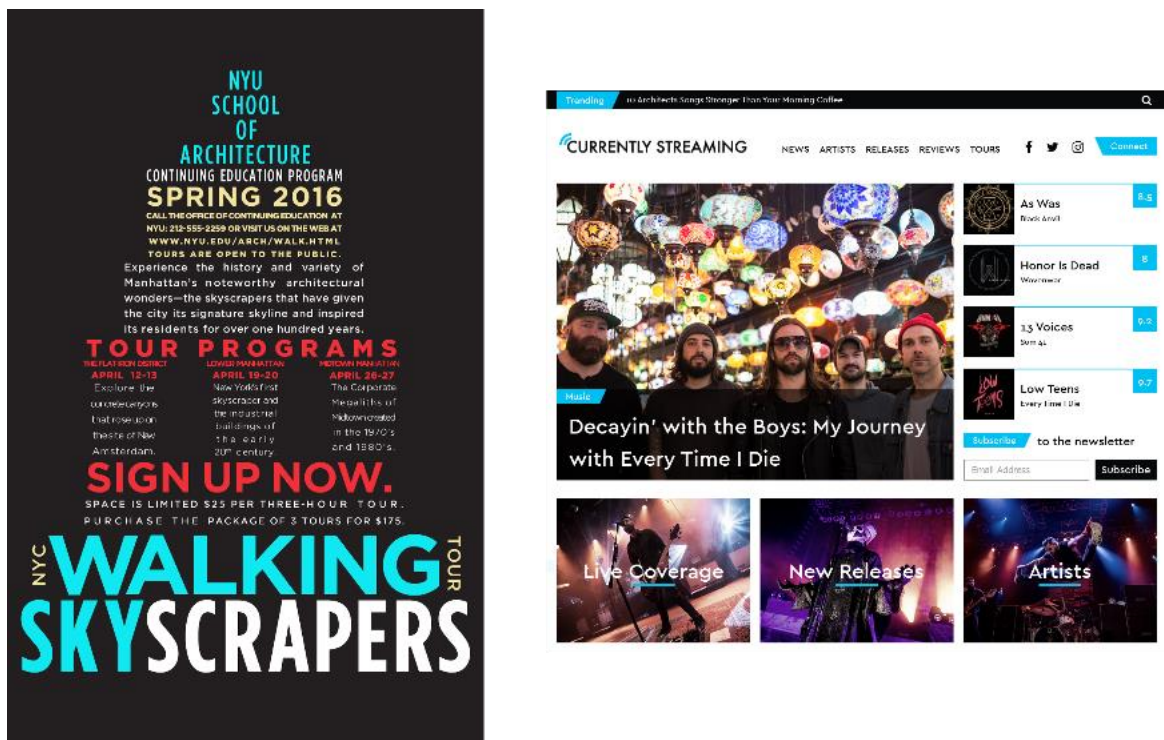


Figure 3.1 The designs used in the study.

3.1.3 Feedback Interfaces

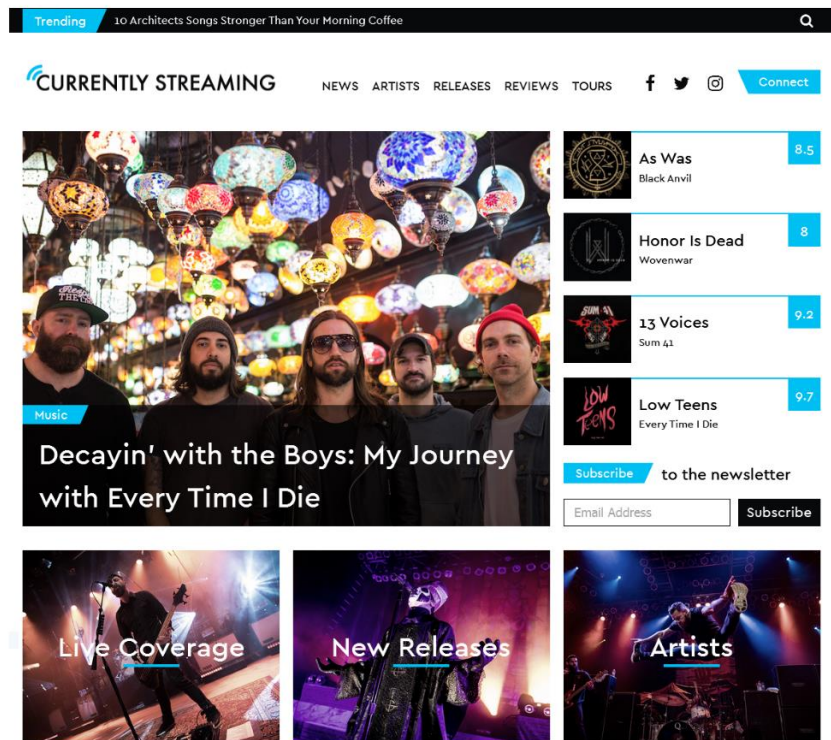
The feedback interfaces consist of text instructions for the feedback provider. Beneath the text is the design to be reviewed. In the text interface, sample feedback is displayed immediately

below the design. Each piece of sample feedback is displayed as a truncated version of the feedback with the text “show more” beside it. Clicking on “show more” reveals the full text of that piece of feedback. This model was adopted to allow us to log which pieces of feedback were viewed by each provider. Under the sample feedback is a text box for providers to enter their comments. Finally, beneath the textbox is a “submit” button for the provider to submit their feedback after they had finished making changes. This interface was designed to be similar to other non-spatial feedback interfaces such as Reddit. The interface is depicted in Figure 3.2.

Your Task

This is an alternative music news website. The target audience is users who are interested in music. The purpose of the website is to be inviting, easy to use, and to connect with and provide relevant articles for people who are passionate about music. Inspect the design for no less than 30 seconds. Feedback submissions that do not demonstrate reasonable effort will be rejected.

Please give feedback on the design and offer any suggestions for improvement.



Here are some examples that other people have provided. Feel free to read as many as you would like.

1. The layout looks like a Wordpress layout.... [Show more](#)
2. Generally, the site looks well-designed and clean.... [Show more](#)
3. The photos are good, intriguing. The... [Show more](#)

Continue

Figure 3.2 The text interface. Providers leave their comments in the text box. In the aesthetic and goal-oriented history conditions, they may choose to view the full text of a piece of sample feedback by selecting “show more.”

In the spatial interface, each piece of feedback is anchored to a point on the design. The provider first selects a position on the design by clicking on it. This places a marker on the position and opens a text box for the provider to input their comments. Providers could place any number of comments in this way. Providers may also select a placed marker to reopen their feedback to review or edit it. In the aesthetic and goal-oriented History conditions, markers representing the sample feedback are placed on the design image when the page loads. Providers may select these markers to review their contents, as above. Figure 3.3 demonstrates the spatial interface.

Your Task

This is an alternative music news website. The target audience is users who are interested in music. The purpose of the website is to be inviting, easy to use, and to connect with and provide relevant articles for people who are passionate about music. Inspect the design for no less than 30 seconds. Please give feedback on 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.

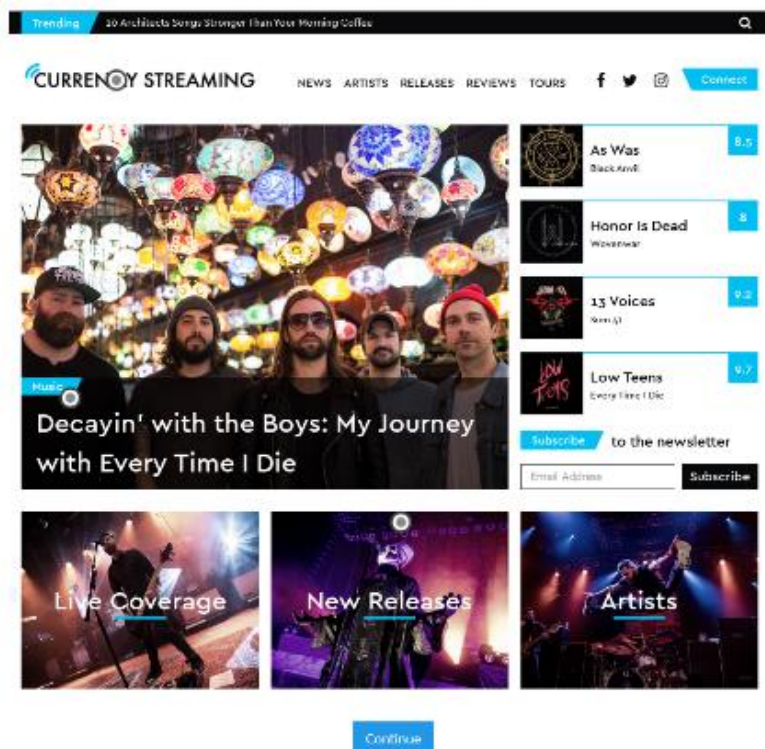


Figure 3.3 The spatial interface. The provider enters feedback by selecting a point on the design and entering text in a window. They could leave any number of comments. They could review the history by clicking on existing markers.

3.1.4 History conditions

There were three conditions of History: no-history, goal-oriented, and aesthetic. Providers in the no-history condition were not able to review feedback. Participants in the goal-oriented condition had the option to review sample feedback which had been judged to address the designers' presumed goals. Participants in the aesthetic condition had the option to review feedback which was not goal-related.

There were three pieces of feedback for the aesthetic history condition and three the goal history condition, as described above. The same feedback was used across Interface conditions. In the spatial interface condition, a marker was placed on the design for each piece of feedback. Providers could view the associated feedback by clicking on a marker. In the text interface condition, each piece of feedback was presented a few characters of the feedback were presented with a "show more" beside it. Clicking on "show more" revealed the remainder of the feedback text.

We decided to use history to explore the area of goal-related feedback, which references the presumed goals of the designer or the as described by Hicks, et al [18]. They found that goal-related statements were correlated with greater perceived quality of design feedback, and that dividing the feedback process into smaller, discrete subtasks led to providers generating more goal-related feedback. We investigate whether access to a history of goal-related feedback would also lead providers to create more goal-oriented feedback themselves, compared to access to history feedback which is not goal-related (aesthetic) or no history.

3.1.5 Sample Feedback

To generate sample feedback for the History conditions, we used the text collection interface in a pilot study to collect feedback for each design. After generating feedback, we asked a new set of providers to rate this feedback on the extent to which the feedback commented on reaching the designer's goals, and the quality of the feedback as assessed by how helpful they thought the feedback would be to the designer. Each measure was assessed on a 5 point Likert scale. Each piece of feedback received three ratings. If the reviewers did not reach agreement by the maximum and minimum scores being within two units for each category, the feedback was discarded. If there was agreement, we used the average of user's scores. We then discarded all feedback whose quality was below the mean. The three pieces of feedback with the highest and the three with the lowest goal-oriented scores were used as the example feedback in the goal-oriented and aesthetic History conditions, respectively. We decided to provide three pieces of feedback in the aesthetic and goal-oriented History conditions because prior studies have found that over 80% of images posted to an online review platform have less than four responses [26].

3.1.6 Procedure

The study was conducted through Mechanical Turk. On accepting the task, the feedback provider was directed to the consent form. If accepted, the subject was randomly assigned to one of the experimental conditions. They were presented with a design and asked to provide feedback through the assigned collection interface. After submitting their feedback, providers were asked to complete a short survey including their expectation for the helpfulness of their feedback, effort required by the task, and demographic information.

3.2 Measures

We measured the content of feedback, perceived quality of feedback, the behavior of providers, and providers' self-assessments.

3.2.1 Content analysis

To analyze the content of provided feedback, we measured the length of feedback and potential fixation on history. We also segmented feedback into individual idea units and rated their category, topic of discussion, and perceived quality.

We measured feedback length as the total character count of all feedback left by a provider.

To examine the impact of viewing sample feedback, we only considered cases where participants examined some but not all the sample feedback. This left us with 56 instances of feedback in the spatial and 68 in the text condition. We then calculated the average similarity scores between the feedback the provider sample feedback the provider reviewed.

We generated the TF-IDF vectors for the feedback and samples. Similarity was calculated as the cosine similarity between the feedback and the example, as implemented in the scikit-learn toolkit for Python [27]. This process has been widely used and is an accepted measure of similarity between documents [19, 28]. Cosine similarity is a measure of text similarity from 0 (low) to 1 (high similarity). For example, the phrase, "an apple a day keeps the doctor away," has a similarity score of 1.0 to itself, but its similarity to, "I would like an apple," is 0.25 [29]. Our hypothesis was that reviewing feedback would induce a fixation effect in providers, and that feedback they leave would be closer to history they have reviewed than to history they have not reviewed.

We partitioned provider feedback into individual idea units. Idea units each represent a coherent thought. Each idea unit was then coded according to the categories of idea units in critique discourse created by [20]. Categories included directive, nondirective, praise, criticism, and other. Directive feedback represents instructions for a specific suggestion for a change to the design. Nondirective represents suggestions which are general enough to be applicable to any design; for example, *“Make it better.”* Praise represents the category of comments which express approval without suggesting changes for further improvements. For example, *“This is great!”* Similarly, criticism represents comments which express a negative evaluation of some aspect of the design, without offering suggestions for how to fix the design. For example, *“I don’t like it.”* We also included the category of investigation from Dannels and Martin (2008) [30], indicating the provider expressed confusion or asked the designer to justify a decision. An example of an investigative feedback would be, *“Why is the word ‘sky’ in blue?”* We included a category of other, if the idea unit’s category was ambiguous or unclear.

In addition to category, we also coded the idea units based on the topic of discussion. Topics included goals, content, color, font, layout, and other. As in Hicks et al [18], goal feedback indicates the feedback explicitly referenced the presumed goals of the designer or the purpose of the design. Color, font, and layout are all forms of aesthetic feedback. Content feedback indicated the subject of discussion was the information directly represented in the text or the design’s functionality (in the case of the website). ‘Other’ was used as a catchall for idea units which were ambiguous or off topic. In total, the feedback was divided into 1045 idea units. Two coders with experience in design categorized and rated the idea units. Reliability was measured as Cohen’s Kappa on a subset of 5% of the dataset. Their reliability ranged from .80-.82 across category, topic, and quality, representing acceptable agreement.

3.2.2 *Perceived quality of feedback*

The quality of each idea unit was rated on a scale of 1 to 5 in accordance with the question, “how helpful would this unit of feedback be to the designer for improving the design?” In this study, each idea unit’s quality was rated independently to account for the potential diversity of specificity and perceived usefulness of feedback. This was done to reach a higher level of granularity in rating feedback. However, prior work has either assigned each category of feedback to a high or low level of quality [4], or else judged the quality of provider’s entire feedback contributions holistically [18, 31, 32].

3.2.3 *Behavioral measures*

We measured the amount of example feedback viewed by providers and the length of time providers took to complete the task.

We calculated the amount of prior feedback examined by providers and computed general behavioral metrics. Provider interactions with example feedback were logged. In the spatial condition, the provider could click on a marker to display the associated sample feedback. In the text condition, the provider could click on “show more” to view the full text of the feedback. We considered the provider as having reviewed any piece of feedback which they had clicked to reveal.

We also recorded the amount of time the provider took to complete the task.

3.2.4 Self-assessment

After completing their feedback, participants were asked to rate the effort required for the task, how useful they thought their feedback would be for the designer, and their level of design expertise. Each question was on a 5 point Likert scale from 1 (low) to 5 (high). They were also asked to provide demographic information (age and gender).

Demographics

Responses to these questions will not affect payment.

Please enter your age:

☐ 18-24

☐ 25-34

☐ 35-44

☐ 45-54

☐ 55 or older

Please select the gender you identify with:

☐ Male

☐ Female

☐ Other

Please rate your design expertise (1: novice, 5: expert):

☐

☐

☐

☐

☐

1

2

3

4

5

How much effort do you think this task required? (1: low effort, 5: high effort)

☐

☐

☐

☐

☐

1

2

3

4

5

Please rate how useful you think the creator of the design will find your feedback. (1: not useful, 5: very useful)

☐

☐

☐

☐

☐

1

2

3

4

5

Get your payment code.

Figure 3.4 The post-task demographics page

CHAPTER 4

RESULTS

Thirty responses were collected per experimental condition, for a total of 360 responses. Responses which were incomprehensible, or whose length of feedback or time taken were outliers (more than twice standard deviations from the mean for their condition) were excluded and replaced.

4.1 Content analysis

4.1.1 Text condition produced longer feedback

Feedback length was measured as character count. An ANOVA detected a main effect of Interface on feedback length ($F(1,358)=14.9$, $p=0.0002$). Length per condition is illustrated in figure 4.1. Tukey's HSD found that the length of feedback in the text condition ($\mu=415$ characters) was significantly longer than that in the spatial condition ($\mu=324$ characters, $p=0.0001$). To analyze whether the effect was due to a greater use of stop words in the text condition, we repeated the analysis with stop words removed, below.

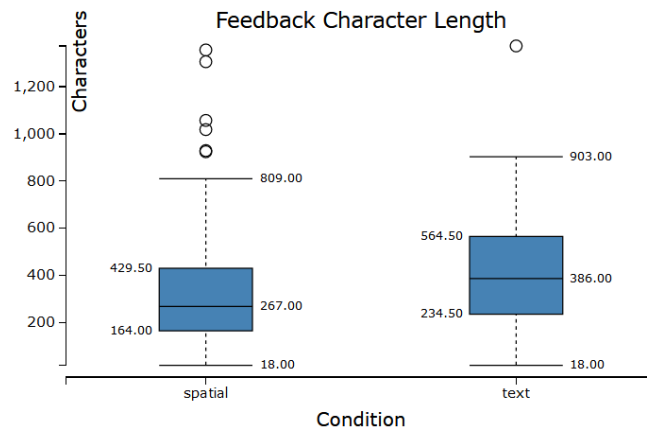


Figure 4.1 Length of feedback. Feedback was longer in the text condition.

4.1.2 Text condition produced longer feedback without stop words

To investigate whether stop words drive the difference in length between conditions, we removed all stop words from the feedback and performed another ANOVA. The ANOVA again found a main effect of Interface on feedback length ($F(1,358)=14.51$, $p=0.0002$). Tukey's HSD also found this difference to be significant ($p=0.0001$). With stop words removed, feedback in the text condition ($\mu=290$) was again longer than feedback in the spatial condition ($\mu=228$). No other effects were discovered.

This effect may be because in the spatial interface, providers do not need to directly refer to or explicitly describe the target of their feedback. For example, one provider in the text condition wrote, "...*The middle of the tower of text has a font that is a bit smaller, and I personally find it hard to read that part.*" In contrast, a provider in the spatial interface wrote a similar statement which read simply, "*This text is too tiny - hard to read.*" Although both samples contain similar information, the provider in the text condition wrote more text. The provider in the spatial condition was able omit any identifying text by leveraging the spatial interface to place their annotation directly on the area of concern on the design.

4.1.3 Generated feedback was more similar to viewed history

To examine the influence of viewing sample feedback, we only considered the contributions of providers who had viewed some (i.e. 1 or 2) but not all the samples. Similarity could then be compared between the providers' feedback and history which they had or had not reviewed. An ANOVA revealed that feedback left by a provider was more similar to the samples the provider viewed ($\mu=0.23$) than to samples the provider did not view ($\mu=0.20$), on a scale from

0 to 1 with 1 being most similar ($F(3,180)=46.72, p<0.0001$). Tukey's HSD showed that the effect was significant ($p<0.0001$). The results are illustrated in figure 4.2.

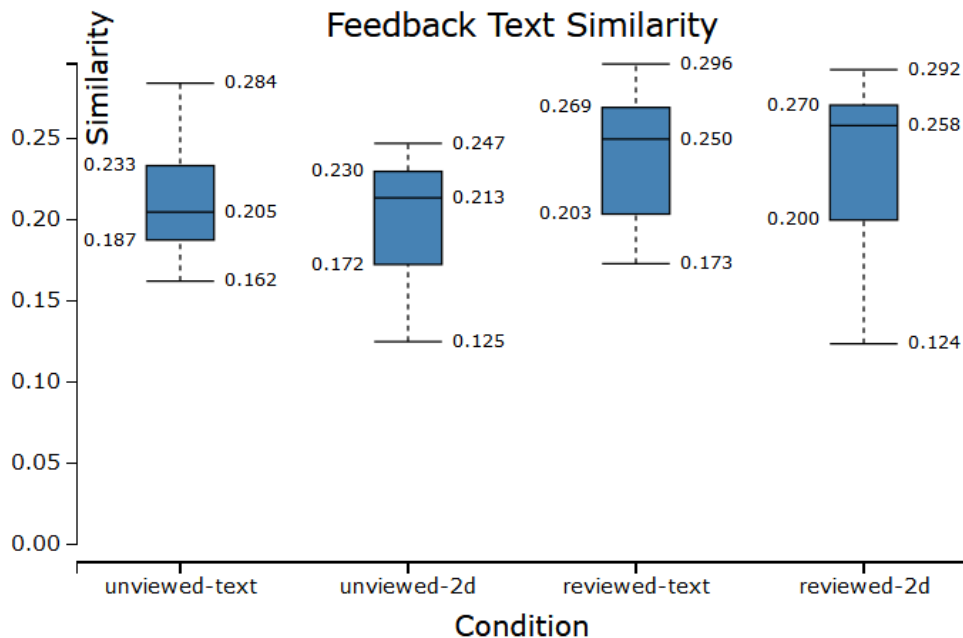


Figure 4.2. Feedback similarity. Providers left feedback which was more similar to sample feedback that provider had reviewed than to feedback the provider did not review.

This implies that the presence of sample feedback or a history of prior comments induces a fixation effect when providing feedback, much as providing pictorial examples introduce a fixation effect in solving design problems. Table 4.1 below displays a piece of sample history feedback and two pieces of feedback which were created by providers, one with a relatively high similarity and one with a low similarity measure. The high similarity feedback uses much of the same language and covers many of the same points as the sample text compared to the low similarity feedback.

	Similarity to sample	Feedback text
History sample feedback	1.0	“The blue is very eye catching, but then the other colors get washed out. The font of the three columns of text are hard to read. I would change the font or try making it bold. There may be too many different colors in the piece. I would get rid of the red and make that text blue or yellow.”
High similarity feedback	0.39	“This text is a little hard to read. I like the red and blue text very much against the black background. I like the design and that it looks like a building but I'd prefer to see it designed without the three columns where the text is hard to read.”
Low similarity feedback	0.11	“I like the symmetry of the overall design and that fact that the fonts are consistently readable, except for this area. I fell [sic] that a bit more spacing will help it be more readable. The colors are excellent and the contrast is eye catching.”

Table 4.1. History feedback and two provider's feedback and their TF-IDF measures of similarity. Providers left feedback which was more similar to history they had reviewed than it was to history which they had not reviewed.

4.1.4 Interface conditions resulted in different categories of feedback

To examine the effect of Interface on the categories of feedback generated, we performed z-tests to find differences in proportions. Table 4.1 contains the frequencies of proportions of Idea Units by History and Interface condition. Providers in the spatial Interface condition left more investigations (9.7%) than in the text Interface condition (5.5%; $z=2.516$, $p=0.012$). They also left more critical comments (20%) compared to the text condition (16.2%; $z=3.06$, $p=0.0012$). In contrast, providers left more nondirective comments in the text condition (1.5%) than in the spatial condition (0.5%; $z=1.99$, $p<0.05$). Those in the text condition also left more praise (40%) compared to providers in the spatial condition (31.5%; $z=2.867$, $p=0.004$).

Investigations are comments which express confusion or ask questions of the designer or the design. Praise and criticism are comments which express positive or negative evaluations but do not offer suggestions for improvement. Nondirective comments are general suggestions which could apply to any design.

	Condition					
	No History		Aesthetic		Goals	
Category	Spatial	Text	Spatial	Text	Spatial	Text
Directive	33.3% (73)	36.2% (55)	32% (58)	36.5% (50)	33.1% (58)	27.1% (49)
Nondirective	0.9% (2)	0% (0)	0% (0)	0.7% (1)	0.6% (1)	3.3% (6)
Praise	30.6% (67)	44.7% (68)	32% (58)	40.1% (55)	32% (56)	3.3% (65)
Criticism	24.2% (53)	12.5% (19)	21.5% (39)	14.6% (20)	14.9% (26)	20.4% (37)
Investigation	9.1% (20)	3.9% (6)	7.2% (13)	4.4% (6)	13.1% (23)	7.7% (14)
Other	1.8% (4)	2.6% (4)	7.2% (13)	3.6% (5)	6.3% (11)	5.5 % (10)

Table 4.2. Frequencies of proportions of idea units by History and Interface condition.

4.1.5 History conditions resulted in different categories of feedback

We used z-tests to examine differences in proportions of categories of idea units as a result of History conditions. We found that providers in the goals condition left more investigation feedback (10.4%) than those in the aesthetic condition (7%; $z=2.075$, $p=0.038$). There were no other significant effects.

4.1.6 Interface conditions caused different feedback topics

We performed z-tests to examine differences in the proportions of topics of discussion of idea units resulting from Interface condition. Spatial providers left more feedback on font (12.2%) than the text condition (8.1%; $z=2.16$, $p=0.0308$). They also left more content feedback (28.3% compared to 13.8%; $z=5.653$, $p<0.0001$). Providers in the text condition left more goal-oriented feedback (13%) than those in the spatial condition (8.2%; $z=2.538$, $p=0.0111$). They also left more feedback on color (18.9% compared to 11.7%; $z=3.287$, $p=0.001$).

Content-related feedback refers to the functionality of the design or the information in the text of the design, without referring to its presentation.

4.1.7 History conditions resulted in different topic categories

We used z-tests to examine differences in proportions of categories provided across different History conditions. The absent History condition resulted in more content-oriented feedback (25.1%) than the aesthetic condition (15.1%; $z=3.24$, $p=0.0012$). The aesthetic History condition resulted in more providers commenting on the color of elements (18.2%) than in the goals History condition (12.1%; $z=2.23$, $p=0.025$). Providers in the goals condition left more goal-oriented feedback (13.2%) compared to those in the aesthetic condition (7.9%; $z=2.24$, $p=0.025$) as well as more content-related feedback (24.4%) than in the aesthetic condition (15.1%; $z=3.03$, $p=0.0025$).

4.2 Behavioral measures

4.2.1 Design influenced amount of sample feedback viewed

The feedback viewed by each provider was logged. An ANOVA found a main effect of Design on the number of pieces of feedback reviewed by providers ($F(1,240)=4.51$, $p=0.03$). Tukey's HSD confirmed that providers in the poster condition reviewed more feedback on average ($\mu=2.0$) than did providers in the website condition ($\mu=1.7$, $p=0.03$).

4.2.2 Providers spent more time generating feedback for the website design

An ANOVA discovered an effect of Design on how much time providers spent completing the task ($F(1,358)=3.94$, $p=0.048$). Tukey's HSD verified this ($p=0.0001$). Providers spent less

time before submitting feedback for the poster design ($\mu=265$ seconds) than they did for the website design ($\mu=309$ seconds).

The providers may have spent longer providing feedback for the website because they were imagining interacting with the site, or it may have had fewer basic design issues or may have been more complex than the poster design.

4.3 Self-assessment

4.3.1 Sample goal-oriented feedback reduced perceived effort

After providing feedback, providers were asked to report how much effort they felt the task required on a scale from 1 (low) to 5 (high). An ANOVA revealed a main effect of History on providers' self-reported effort scores ($F(2,358)=3.65, p=0.03$). Effort ratings were higher for the no-history condition ($\mu=3.5$) than for the goal-oriented sample feedback ($\mu=3.1$). Tukey's HSD found a significant difference to be significant ($p=0.02$). This suggests that providers may find a task easier not only when there is sample feedback, but when the sample feedback also references the designer's presumed objectives.

No other significant results or interactions were observed.

CHAPTER 5

DISCUSSION

This paper examines how collection tool interface and the presence and nature of history can influence feedback. Differences in interface and the presence or absence of history caused significant changes in the content of reviews. We found that feedback left by providers was longer in the text condition than for the spatial condition. Further analysis showed that this was not due to stop words. It is possible that the providers in the spatial condition could leave less feedback while providing a similar amount of information by leveraging the fact that they could place an annotation directly onto the design. In contrast, a provider in the text condition who wanted to reference a piece of the design would need to describe that part. Reviewers using a spatial interface are able markup targets directly rather than needing to describe them, and this may account for the difference in length. Although feedback in the text condition was longer, it is not clear whether that feedback is more substantive; we did not find a significant difference in the quality of feedback generated by the text and spatial conditions.

Many feedback platforms allow reviewers to see the comments that have been left by prior reviewers before generating their own feedback. We found that having a history of feedback available induces a fixation effect in reviewers across the categories of history ($p < 0.0001$), causing providers' own feedback to be more similar to other feedback which they had reviewed.

In the design process, both divergent and convergent processes must be employed. Divergent feedback helps designers to generate and consider alternatives, and convergent thinking helps designers to select which alternatives they will use to proceed [33]. Divergent feedback is most useful earlier in the design process, and convergence is more useful in later phases. Designers may wish to permit reviewers to see prior feedback on their designs if they would prefer convergent

feedback. In this case, the designer may choose a platform where prior feedback is visible to reviewers. Greenberg et al [4] suggested that designers find consensus within crowd feedback to be particularly useful. However, fixation can also be dangerous because providers may fixate on and recreate counter effective elements.

If the purpose of gathering feedback online is to obtain divergent, independent feedback, it is better for designers to choose to not allow reviewers to see each other's feedback. Divergent feedback may be particularly useful at earlier stages of design [4, 33]. Hicks et al suggested [18] that interfaces which require open-ended feedback such as those discussed here may also lead providers to give more developmental rather than evaluative feedback. This would also suggest that the types of interfaces investigated here would be more useful early in the design process.

An exception in the case of platforms which allow feedback providers to respond to one another's comments would be the possibility of creating a discussion, which could result in deeper, more thoughtful feedback. This may result in higher quality feedback in some circumstances, but extended discussion is not the norm in online feedback platforms [26]. Although feedback is more similar to the sample feedback viewed by the provider, it is unclear to what extent reviewers are inspired to examine those points for themselves or are merely echoing the content of prior feedback.

However, because the difference in similarity to viewed feedback was small (.23 compared to .20), this suggests that providers do not simply copy prior feedback directly. A more detailed analysis will be necessary to examine the full impact of exposure to prior feedback has on the quality of new contributions and how to encourage fruitful discussion.

We also found that the feedback collection interface influences the content of the feedback providers generate. There were several differences in both the category of idea units and topic of

discussion with experimental condition. Providers in the text interface condition appeared to leave more positive comments overall, with more praising comments (40%) than the aesthetic condition (31.5%). Similarly, providers in the text condition left fewer critical statements (16.2%) than those in the aesthetic condition (20%). This may be due in part to the fact that providers in the spatial condition can easily view all of their feedback together, whereas in the spatial condition feedback is fragmented across multiple annotations. Providers in the text condition may be conscious of the overall tone of their feedback and feel social pressure to tend adjust their feedback towards being more positive.

Providers in the spatial interface left more investigation feedback (9.7%) than those in the text condition (5.5%). Investigation involves asking questions of the designer, and these comments may serve to highlight areas of potential confusion for the audience. Because the spatial condition involves visually scanning the design, providers in that condition may have been more likely to discover less salient parts of the design and to consider these without as much of their context. Both could lead providers to ask more questions.

Providers also left nondirective or general suggestions, although the number of these was low for both the text (1.5%) and spatial conditions (0.5%). These are suggestions which are so general that they are unlikely to be useful. One reason why they may have been more common in the text condition is that these providers must consider the design as a whole when generating feedback, which may have led them to make more general statements. In contrast, when a provider responds to part of a design, they are more likely to have a specific comment or complaint to address.

There were differences in the topic of discussion between feedback from the interface and history conditions. Providers in the spatial condition were more likely to leave feedback on the

font of the design (12.2%) than those in the goals condition (8.1%). Similarly, the spatial interface also led to more feedback on color (18.9%) than the text interface (11.7%). Taken together, this suggests that providers in the spatial condition do consider more aesthetic or surface-level aspects of a design. As we expected, providers in the text condition left more goal-oriented feedback (13.2%) compared to the spatial condition (7.9%). These findings indicate that the interface for feedback collection and the presence of example feedback may lead providers to comment on certain aspects of a design. Designers might then choose the interface which tends to induce the kind of feedback they are most interested in.

Those in the spatial condition were also more likely to comment on the content presented rather than the design itself (28.3%) than were those in the text condition (13%). This includes questions, comments, and suggestions for what information should be presented or functionality for the website. For example, several providers suggested displaying rotating links to articles in a banner near the top of the site. However, such feedback may not directly pertain to the question of how to best improve the current design. Providers in the spatial conditions may be particularly likely to discuss the content of the design because visually scanning through the design may lead them to question the functionality of particular points or to brainstorm ideas for parts which could be added to the design but which may be redundant or irrelevant. Scanning through the design elements visually may lead providers to assess elements individually expect them to stand alone.

Providers in the goals history condition also left significantly more investigation feedback than those in the aesthetic history condition (10.4% and 7%, respectively). Providers who had reviewed comments which referenced the designer's goals may have been inspired to question how and whether the design meets its goals. Similarly, providers in the goals history condition were more likely to leave goal-oriented feedback (13.2%) than those in the aesthetics history

condition (7.9%). Providers in the aesthetic history condition were more likely to leave feedback about color (18.2%) than those in the goals history condition (12.1%). This indicates that providers are more likely to comment on matters related to topics in reviewed history, and may be due to providers becoming fixated on comments which they review.

There was no significant difference observed between the amount of feedback reviewed by providers across the goal and aesthetic history conditions ($\mu=1.8$). Moreover, neither interface condition saw providers examining more sample feedback than the other. This is in line with expectations of how much feedback providers would review. This also indicates that either interface may be equally useful for allowing providers access to historical feedback. This study limited the amount of feedback to three pieces to reflect situations common at the time this study was conducted. However, these results may not extrapolate to cases where designers wish to present reviewers with many pieces of feedback.

Providers took significantly more time providing feedback for the website design than for the poster ($p=0.03$). Websites, unlike posters, are an inherently interactive medium. Although they were presented with a screenshot, providers may have spent time imagining how they interact with the site to consider ease of use. Since providers gave more similar feedback to the poster design, it is possible that the poster had more basic design issues to address. If the website has fewer basic issues, providers may have needed more time to identify them. Alternatively, providers may have perceived the website design as being more complex and thus taken more time to study it before leaving feedback. Further work is needed to explore how reviewers organize their time considering documents and structuring their responses.

Self-reported measures of effort (“how much effort do you think this task required?” rated on a scale of 1-5) were collected from each provider at the end of the study. Analysis found that

the self-reported effort was lower for providers who could view goal-oriented feedback (3.1) than for providers who were not able to view sample feedback (3.5, $p=0.02$). There was no significant difference observed between aesthetic feedback and other History conditions. This suggests that if designers provide sample feedback which addresses their own goals, this may aid reviewers in generating their own feedback [10]. This may work by better framing the context in which the feedback is needed and how it might be used. However, the providers in the goal-oriented feedback condition did not significantly assess their feedback as being higher quality in terms of being more useful to the designer.

5.1 Limitations

AMT providers have financial motives. They are incentivized to satisfactorily complete as many tasks as possible within a given time. These motives set them apart from other groups who commonly give or need design feedback, including students and professional designers. Students may be driven by curiosity, social factors, and a desire for grades. Experts may be motivated by their professional reputation. Since these populations have different motives more approaching design feedback, we do not know if the findings here will apply directly to these other contexts.

We also only investigated two interfaces, a spatial interface which allowed users to leave feedback by placing annotations on the image itself, and a text interface in which users leave all feedback in a text box below the image. Review platforms such as CrowdCrit and Voyant mix spatial and text modalities, allowing reviewers to both annotate a design and to enter comments in a free form text box. These interfaces with multiple modes of feedback may elicit different behavior than those with a single mode alone.

This study also only considered two designs for providers to generate feedback. Different designs may afford different kinds of feedback depending on what issues are most salient. Although this study found differences in the idea unit categories and topics of discussion, it is possible that another study using different designs would have different findings. The differences found in the discussion of may be due to the designs used, rather than aspects of the interfaces used.

We did not consider what effect the expertise of reviewers may have had. For example, experts may find it unnecessary to review sample feedback, whereas novices may find inspiration from examples [10, 32]. AMT providers represent a diverse population from different backgrounds and varying levels of experience. Prior work has contrasted the feedback provided by AMT providers and that of experts [10]. Other approaches have required providers to demonstrate a level domain knowledge before being allowed to accept a task [2]. A future study which assesses the quality of feedback may also investigate whether there is a correlation between the quality of feedback and the provider's level of expertise.

5.2 Conclusion

Designers are quickly adopting a range of online tools and communities to collect feedback on their work. The focus of this study was investigating how the feedback collection interface and available history of feedback influence the generation of new feedback. In the conceptual stages of design, the results from this study suggest that a designer should choose feedback collection platforms which do not share feedback between providers. Later in the design process, designers may wish to gather more convergent evidence under some circumstances. In this case, the designer may wish to select a feedback collection platform which allows users to review feedback from

other providers, such as Dribbble or Reddit. This study finds that designers who wish for suggestions which consider their goals and the purpose of the design should use a collection tool with a text interface. Designers who wish for feedback on the color or font of their design would be best served by using a spatial feedback collection interface. Designers who prefer comments on what content should be included in their design should also use a spatial interface. If feedback is made available to other providers, it will influence the feedback which they generate. Feedback collection platforms should allow designers to choose whether providers can review prior feedback, and provide guidelines on when a public feedback history is advisable. We hope this work contributes to the understanding of extant tools, aids in the development of new interfaces, and helps designers in selecting the best feedback collection tools for their needs.

REFERENCES

1. Elkins, J., *Art critiques : a guide*. 2nd ed. 2012, Washington, DC: New Academia Publishing. xiii, 215 pages.
2. 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 & social computing*. 2014, ACM: Baltimore, Maryland, USA. p. 1433-1444.
3. 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 & Social Computing*. 2015, ACM: Vancouver, BC, Canada. p. 1637-1648.
4. 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.
5. *Reddit*. Available from: https://www.reddit.com/r/design_critiques.
6. *Dribbble*. Available from: <https://dribbble.com>.
7. *Adobe Acrobat*. Available from: <https://get.adobe.com/reader/>.
8. *Red Pen*. Available from: <https://redpen.io>.
9. Luther, K., et al., *CrowdCrit: crowdsourcing and aggregating visual design critique*, in *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, ACM: Baltimore, Maryland, USA. p. 21-24.
10. Luther, K., et al., *Structuring, Aggregating, and Evaluating Crowdsourced Design Critique*, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 2015, ACM: Vancouver, BC, Canada. p. 473-485.
11. Jansson, D.G. and S.M. Smith, *Design fixation*. *Design studies*, 1991. **12**(1): p. 3-11.
12. Chrysikou, E.G. and R.W. Weisberg, *Following the wrong footsteps: fixation effects of pictorial examples in a design problem-solving task*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2005. **31**(5): p. 1134.
13. 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.
14. 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: Montréal, Québec, Canada. p. 1243-1252.
15. 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.

16. Gero, J.S., *Fixation and commitment while designing and its measurement*. The Journal of Creative Behavior, 2011. **45**(2): p. 108-115.
17. Viswanathan, V. and J. Linsey. *Understanding fixation: A study on the role of expertise*. in *DS 68-7: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 7: Human Behaviour in Design, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011*. 2011.
18. 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.
19. Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Vol. 1. 2008: Cambridge university press Cambridge.
20. Cho, K., C.D. Schunn, and D. Charney, *Commenting on Writing*. Written Communication, 2006. **23**(3): p. 260-294.
21. Novak, E., R. Razzouk, and T.E. Johnson, *The educational use of social annotation tools in higher education: A literature review*. The Internet and Higher Education, 2012. **15**(1): p. 39-49.
22. Zyto, S., et al. *Successful classroom deployment of a social document annotation system*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012. ACM.
23. Glassman, E.L., et al. *Mudslide: A spatially anchored census of student confusion for online lecture videos*. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015. ACM.
24. 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.
25. Yen, Y.-C., et al., *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*. 2016, ACM: Brisbane, QLD, Australia. p. 773-784.
26. Xu, A. and B. Bailey, *What do you think?: a case study of benefit, expectation, and interaction in a large online critique community*, in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 2012, ACM: Seattle, Washington, USA. p. 295-304.
27. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 2011. **12**(Oct): p. 2825-2830.
28. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval*. Information processing & management, 1988. **24**(5): p. 513-523.
29. *Similarity between two text documents*. 2012; Available from: <https://stackoverflow.com/questions/8897593/similarity-between-two-text-documents>.

30. Dannels, D.P. and K.N. Martin, *Critiquing critiques a genre analysis of feedback across novice to expert design studios*. Journal of Business and Technical Communication, 2008. **22**(2): p. 135-159.
31. Hui, J., et al. *Using anonymity and communal efforts to improve quality of crowdsourced feedback*. in *Third AAAI Conference on Human Computation and Crowdsourcing*. 2015.
32. 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.
33. Daly, S.R. and S. Yilmaz, *Directing Convergent and Divergent Activity through Design Feedback*. Analyzing Design Review Conversation, 2015. **21**: p. 413-429.