Text and Visual Annotation Tools for Scalable Design Feedback Generation

| 1st Author Name  Affiliation  City, Country  e-mail address | 2nd Author Name  Affiliation  City, Country  e-mail address | 3rd Author Name  Affiliation  City, Country  e-mail address |
| --- | --- | --- |

# ABSTRACT

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

* ACM copyright: ACM holds the copyright on the work. This is the historical approach.
* License: The author(s) retain copyright, but ACM receives an exclusive publication license.
* Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Times New Roman 8-point font. Please do not change or modify the size of this text box.

Each submission will be assigned a DOI string to be included here.

Use of crowd feedback systems has been shown to lead to improved visual designs. But there is little research on the relationship between characteristics of the feedback collection interface and nature of the feedback collected. … results …

## Author Keywords

Crowdsourcing; design; feedback; creativity.

## ACM Classification Keywords

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

# INTRODUCTION

Collecting and addressing feedback are vital steps in the iterative design process. Generated insights help guide designers towards artifacts that better connect with the target audience [1]. Soliciting design feedback online offers several advantages over alternatives due to enhanced scalability, availability, and affordability [2, 3]. Designers have many different classes of platforms to choose from when soliciting feedback online. One way of organizing these classes by type of interface they offer to the feedback provider.

One class of these tools is a spatial interface where the feedback provider must first visually mark a location on the design before they are able to enter feedback. Examples of the spatial interface class include tools such as Redpen.io and Adobe Acrobat [4, 5]. Another class of these tools uses a text-centric open ended interface. Interfaces in this class such as Reddit and Dribble do not require the provider to first visually mark the design [6, 7].

But small differences in the feedback generation interface can significantly affect the provider’s behavior. For instance, it has been shown that splitting design feedback task into multiple shorter tasks elicits more diverse feedback [8]. Researchers have employed techniques such as scaffolding resulting in enhanced quality of collected feedback [9, 10]. Asking a feedback provider to spatially mark the design requires them to visually search the design and focus their attention on specific details [3, 11]. Whether the design tool reveals history feedback could also influence a provider’s behavior. Revealing previous feedback to a provider, while encouraging novel ideas, may also encourage conformity and introduce a framing effect [12, 13]. It is therefore important for designers and interface developers to understand the relationship between characteristics of the interface and the resulting feedback for successfully facilitating feedback exchange.

One challenge is that there is little to no empirical evidence that could guide a designer’s decisions or implementations in existing platforms. For instance, difference classes of interface might not generate the same distribution of type of feedback. It is unknown whether the presence of a history enhances creativity or introduces a fixation effect.

In this paper, we present 2 interfaces for soliciting design feedback and study characteristics of the collected feedback in each condition. Interface conditions reflect real-world feedback collection applications such as Reddit and Red Pen [4, 6, 7]. We also manipulate presence of history to generate conditions. Whether to use an interface that presents history is another decision that designers must make when choosing which tool to use for soliciting feedback [14].

We then recruit human participants to provide feedback on visual designs using one of the conditions. We investigate how choosing a non-spatial or spatial feedback interface and hiding or revealing history impacts feedback content. This relationship between provider interface and generated feedback is explored across a variety of visual designs (poster, static website, and web interface).

Among a number of results, we found that presence of history introduces a fixation effect. Feedback generated using the non-spatial class of interface was longer and had more stop words. We also found that classes of interface produce different categories of feedback.

A key contribution of our study is introducing an empirical body of knowledge about ways aspects of the feedback collection interface influence generated content and provider behavior. We believe this knowledge will contribute to building more extensive interfaces feedback exchange, help providers better communicate their evaluation of the design, and help designers improve their work through higher quality feedback.

# Related work

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

## Feedback Collection Tools

There are at least four classes of online tools for collecting design feedback and conducting peer review. One class is spatial annotation tools. In these tools, a feedback provider must first select a region of the design to enter feedback. Adobe Acrobat and Redpen.io implement this class of tool [4, 5]. One advantage of these tools is that requiring feedback providers to visually search for and mark features can focus their attention on those details [15]. But it may also introduce a fixation effect, introducing an inability to see new ways of problem solving [16].

Another class is non-spatial. With the non-spatial tool, feedback providers can see an image of the design and can enter their feedback into a textbox. Reddit and Dribble implement this class of tool [6, 7]. These tools may encourage the feedback provider to generate longer feedback and more conceptual feedback because they don’t have to attach their comment to a specific element in the design. But this less actively engaging interface paradigm could reduce the diversity of generated feedback [8].

A third class of tool for the more general class of peer review is multi-modal. These tools track pen hovering movements in tandem with voice and digitizer writing. An implementation of this tool is RichReview++ [17]. Students preferred using a multi-modal interface to meeting in person. But the linear and irreversible nature of voice makes the commenting task more stressful since students had to think and speak at the same time.

A fourth class is visual. In a visual tool, providers use an image browser to compile their feedback. Moodsource is an example of a visual tool for collecting design feedback [18]. These tools are especially useful for communicating first impressions. But limiting providers to using images may make him unable to convey their ideas.

Our work targeted two classes that are widely utilized for giving design feedback in an online context, spatial and non-spatial. In addition, we also studied these two classes of tools with the presence of history.

For each class of these tools, the interface can choose whether or not to reveal history feedback. Applications such as Adobe Acrobat and email are examples of tools that refrain from displaying history feedback to providers. Choosing to conceal history feedback can reduce the impact of fixation [16] but may also reduce the creativity of feedback providers [14, 19]

However, there is no prior work on how these different classes of feedback collection tools or the presence of history influence the generated feedback. Our work addresses this gap.

## Studies of Crowd Feedback Systems

A number of studies have been conducted studying the benefit of crowd feedback in the design process. The use of crowd feedback systems has been shown to lead to improvements in designs [2]. Systems such as Voyant and CrowdCrit have shown that useful feedback can be generated even from non-expert crowds [3, 9]. Critiki introduced a system that simplified the process of creating, distributing, and aggregating crowdsourced design critique [10].

Most of these systems focus on the interface for the designer. For instance, in Voyant, a bi-directional interaction to link overviews of feedback content and annotations overlaid on the design [3]. CrowdCrit showed the distributions of responses from the providers in order to extract the highest priority issues for the designer who is getting the feedback [9]. Our work seeks to expand on this body by focusing on the experience of the feedback provider.

Studies that focus on the feedback provider have looked at techniques such as rubrics and scaffolding to enhance feedback quality. Critiki found that walking the provider through the critique produces more high quality feedback [10]. In other domains such as peer feedback, framing task goals has been found to significantly change reviews [8]. We are adding to this thread of research by comparing two additional ways of collecting feedback, using a spatial and non-spatial interface.

**METHODOLOGY**

Our study compared how two classes of Feedback Interface (spatial and non-spatial) and History (absent or present) influence generated feedback. We seek to answer the following questions concerning these aspects of the interface: whether they cause the provided feedback to be more specific or more 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.

## 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.

## Participants

Feedback providers (N=360) were recruited from Mechanical Turk. To participate, providers were required to have successfully completed at least 50 tasks and to have a task approval rate greater than 95%. **Insert demographic** **data here**. Based upon a pilot study, they payment was set at $0.50 per task to reflect current US minimum wage.

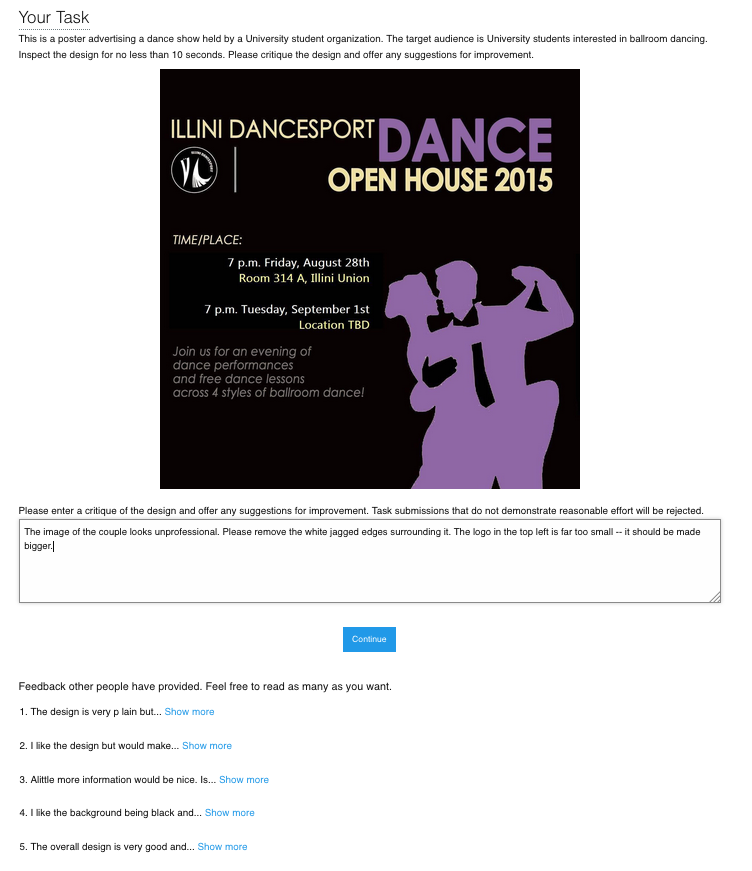
## Designs

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.

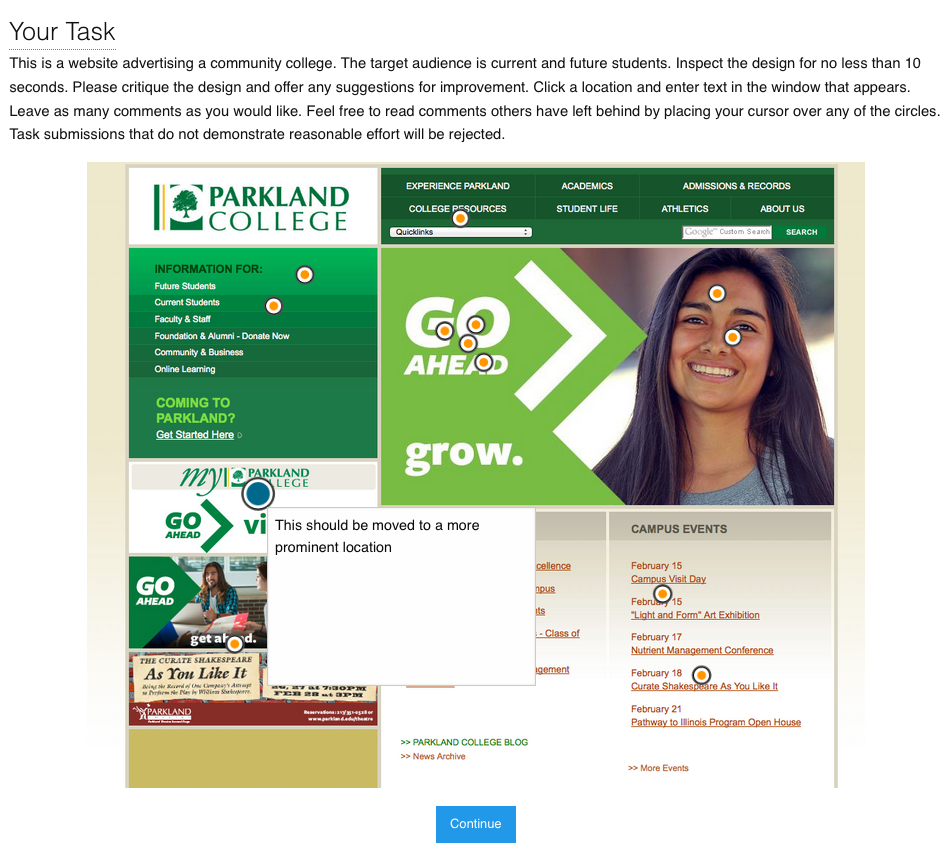
## Feedback Interfaces

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 1 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. In the presence of the history, past feedback was displayed underneath this form. Rather than pre-generating history, we mimicked real world systems by allowing the history to grow organically from feedback submitted by previous providers.

****

**Figure 1. 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 2. 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 desired and were allowed to look at the history by hovering existing markers.**

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 [6, 7]. We adapted this format however to include a “Show more” interaction which allowed us to log which pieces of feedback were viewed.

Figure 2 introduces the spatial Interface. Here, 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. A visual marker representing the feedback is then overlaid on the design to represent the feedback left at the location. The provider could leave as many pieces of feedback as desired, inspect the feedback they had left by hovering over the associated visual marker, and could always edit their own feedback by clicking a marker. In the presence of the History, the pieces of feedback left by previous providers were shown. The participant was allowed to hover over any visual markers to reveal the annotated feedback. Otherwise, the interface operated the same as the History absent condition. The spatial condition was designed and implemented to reflect popular annotation feedback tools such as Adobe Acrobat and Red Pen [4, 5]. Once satisfied with the feedback, the provider submitted their work.

## Procedure

Upon accepting the task, the feedback provider was presented with a consent form. If accepted, they were randomly assigned to one of 12 experimental conditions. The experimental 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.

## Measures

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

*Content analysis*

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

A measure of specificity was calculated for each feedback response. 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”). In calculating sentence specificity, 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].

To categorize the feedback content, 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 categorized each idea unit according to the taxonomy. 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 text length by cumulative character length of all feedback from a single provider.

*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.

Next, for each provider, we aggregated the set of prior feedback that they viewed. For each comment that a provider left, we aggregated the set of feedback they had viewed up to that point. We computed the distances between the recent comment and history feedback using the cosine similarity similarity metric as implemented in the Python pattern.en toolkit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **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 1. Frequencies of the categories of idea units by Interface and History** |

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

*Self-assessment*

Following the feedback task, the provider completed a self-assessment survey. Providers rated their design expertise, perceived effort, and perceived usefulness of the feedback given 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).

# 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. 3 submissions were excluded, leaving us with 357 feedback responses of reasonable quality.

**Content Analysis**

*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 2. 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.

*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 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). An *investigation* is when a feedback provider asked questions about a specific piece of the design. Requiring the feedback provider to select a location on the design before entering feedback may cause them to visually scan the elements in the design before performing feedback entry.

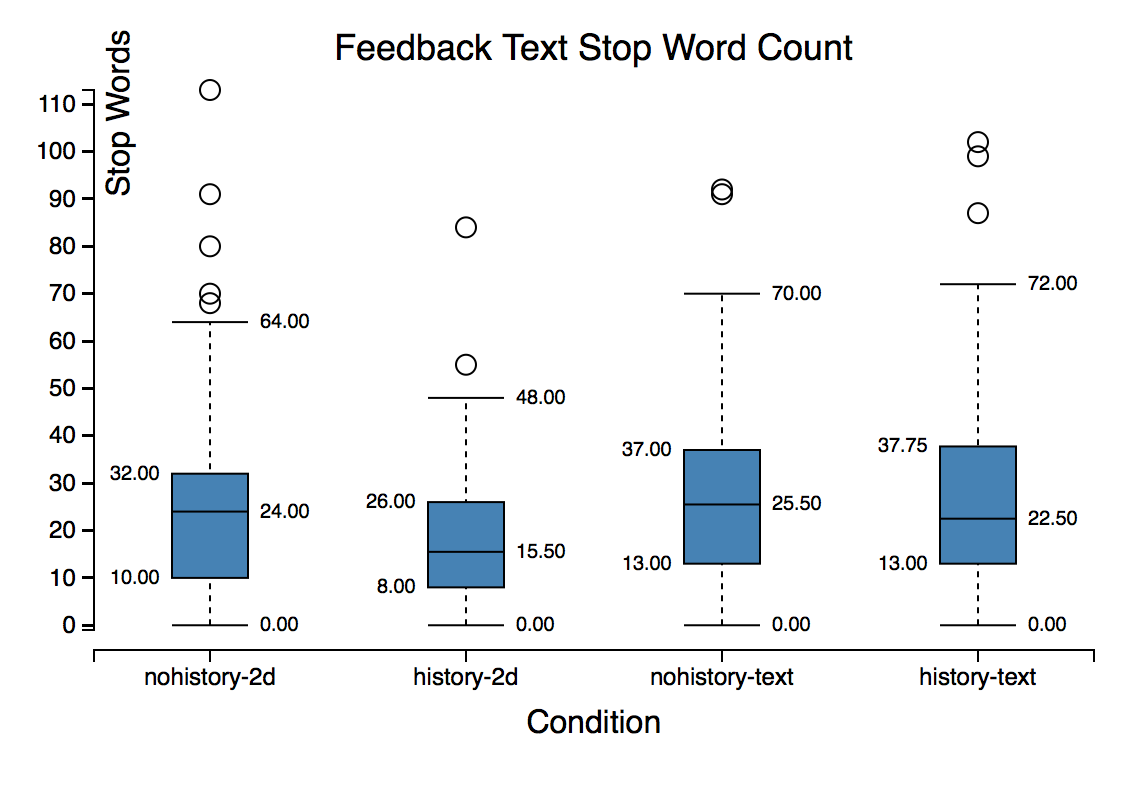
No other significant results were discovered.

*Non-spatial feedback had more stop words*

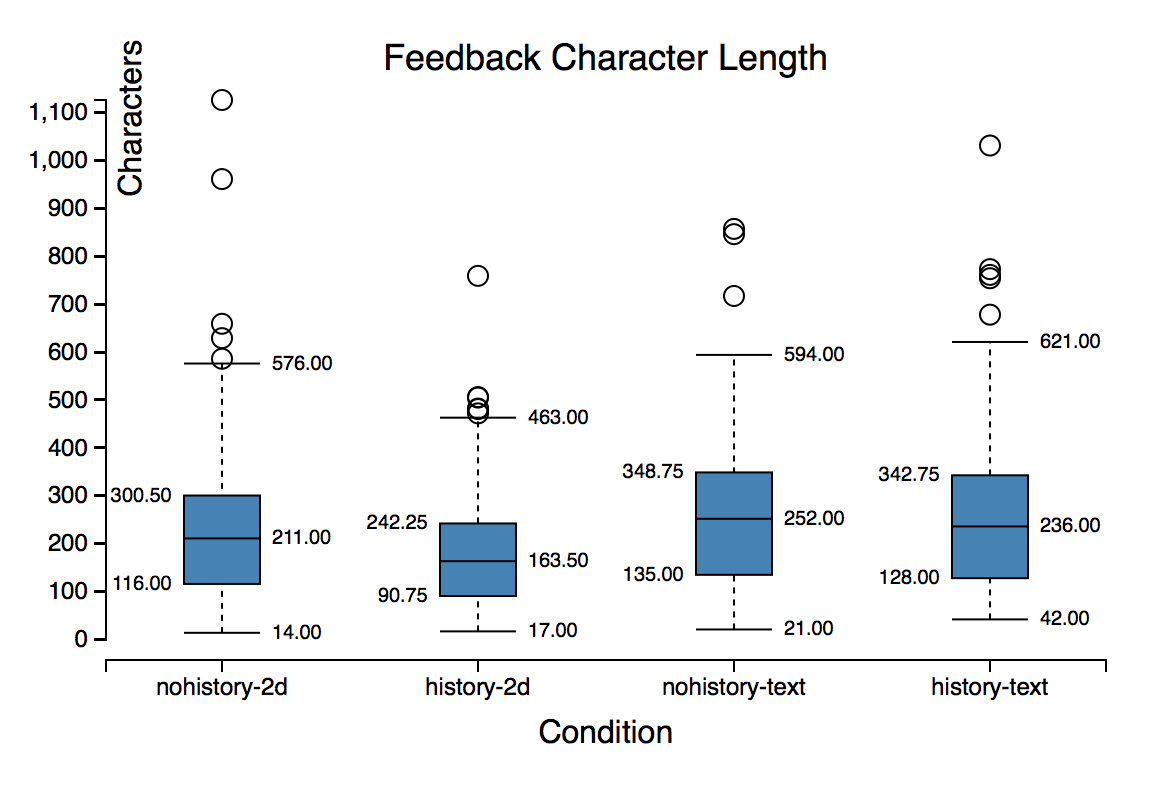
An ANOVA did not detect a main effect of Interface or History on feedback specificity. In the spatial condition, mean specificity was 0.34 (*σ* = 0.17), while the non-spatial condition had a mean specificity of 0.37 (*σ* = 0.14).

An ANOVA uncovered a main effect of Interface on stop word count (F(3,357)=6.93; *p*=0.0089). Figure 1 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).

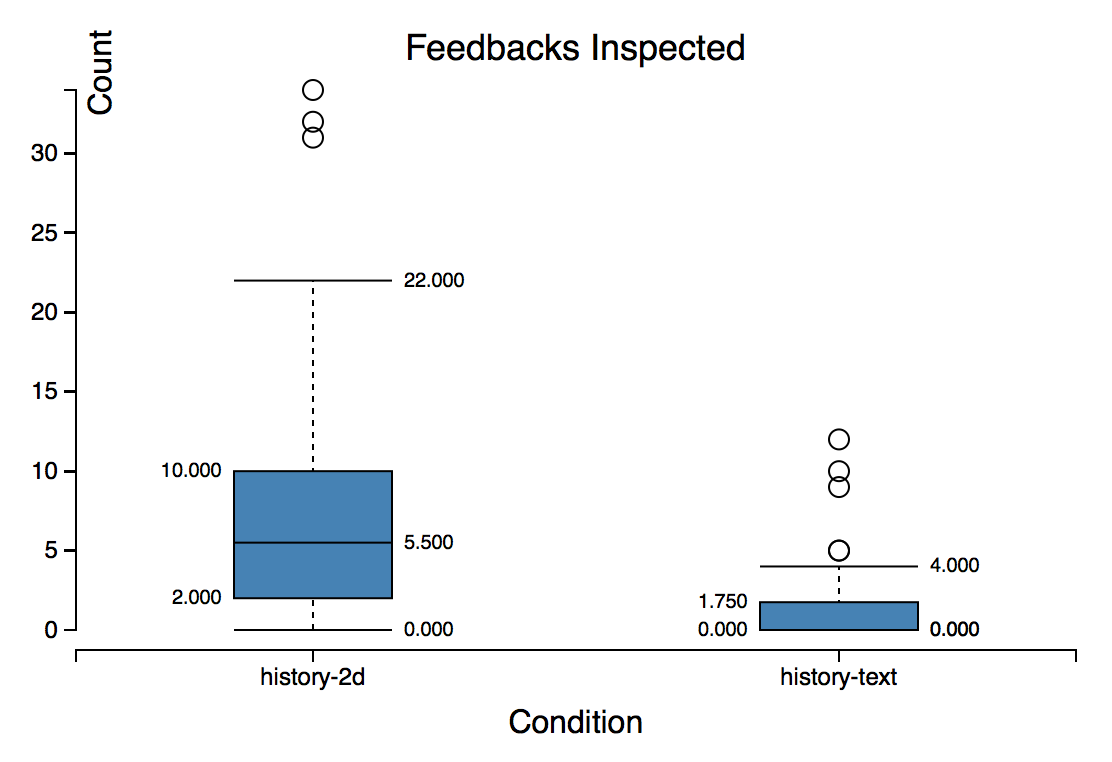
One explanation for this is that the context provided by the spatial condition reduced the need for language necessary to convey the same information in the non-spatial condition. In the non-spatial condition, stop words were used to reference specific elements of the design: *“The logo must come at 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.”*



**Figure 1. This chart shows how the experimental condition affected stop word count of the feedback content. Analysis show providers included more stop words in their feedback in the non-spatial condition.**

**

**Figure 2. 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.**

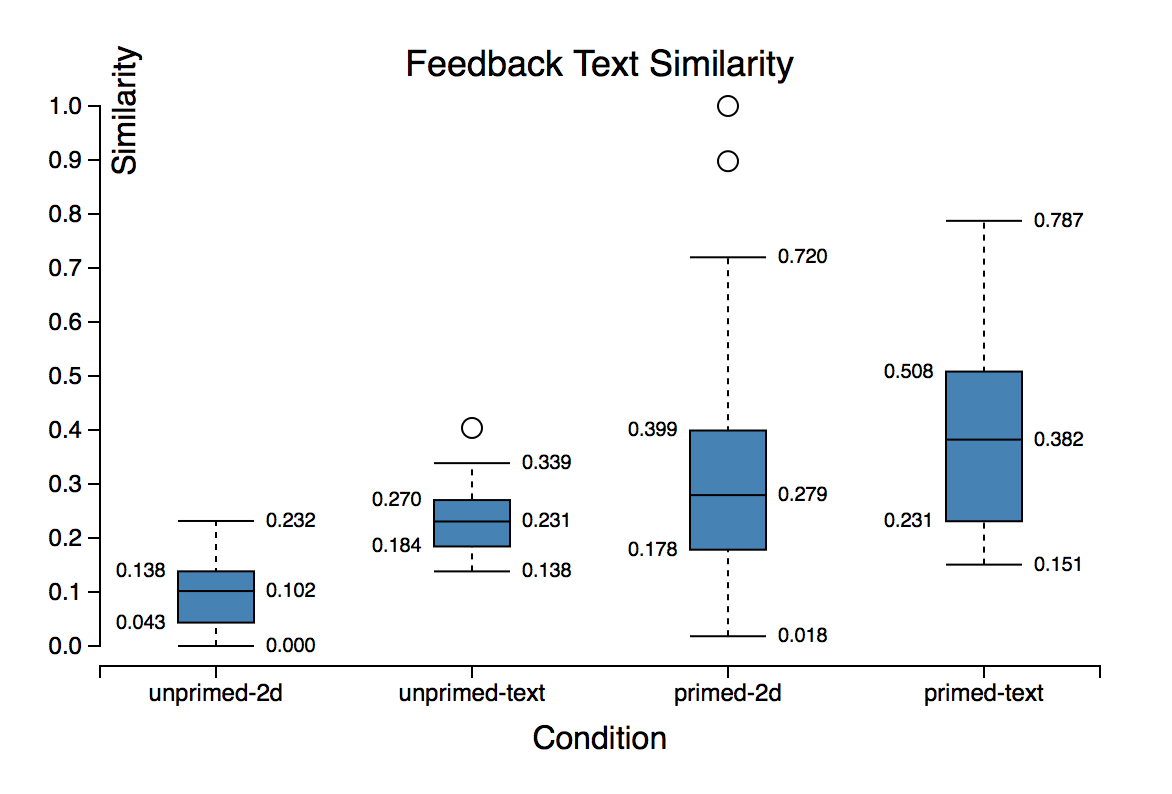
**

**Figure 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.**

**Behavioral Measures**

*Providers inspect more feedback in spatial Interface*

When History was presented in the spatial Interface, we found 55% of providers (99 providers) inspected history feedback. When History was presented in the non-spatial Interface, we found 19% of providers (33 providers) inspected history feedback. The number of instances of feedback inspected by providers is visualized in Figure 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. Similarity scores of generated feedbacks compared to viewed and unviewed history by condition. Generated feedback was more similar to viewed history.**

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

*Generated feedback was more similar to viewed history*

We only considered instances of generated feedback where the provider had examined the 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 displays feedback similarity scores.

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

This suggests that presence of a History introduces a fixation effect for the feedback providers. This effect is analogous to how pictorial representations of examples introduce a fixation effect when solving design problems [16].

*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 viewed feedbacks 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 3.

This effect may have been due to the non-permanence of the feedback window in the spatial Interface. This interface required the feedback provider to hover over a marker to reveal the content. This was in contrast to the non-spatial Interface where an inspected feedback remained visible until the provider explicitly chose to hide it.

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).

|  |  |  |
| --- | --- | --- |
| **Effort Self-assessment** | | |
|  | Non-spatial | **Spatial** |
| History Absent | *μ*=3.3; *σ*=1.1 | *μ*=3.1; *σ*=1.2 |
| History Present | *μ*=3.3; *σ*=1.2 | *μ*=3.1; *σ*=1.0 |

**Table 2. Provider perceived effort self-assessment by condition. Conditions had no significant effect of perceived effort.**

**Self-Assessment**

*Design influenced perceived usefulness of the feedback*

Table 2 shows the breakdown of effort ratings across conditions. ANOVA did not detect differences between these conditions.

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 (*μ*=4.1; *σ*=0.86) and Design C (*μ*=4.0; *σ*=0.89) was higher on average than that of Design A (*μ*=3.8; *σ*=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.

## DISCUSSION AND FUTURE WORK

The goal of our work was to study the influence of Interface and History on generated feedback. We found that the presence of a History may introduce a fixation effect. This fixation effect causes feedback providers to generate feedback that tends to be more similar to the history feedback they inspected. While we weren’t able to distinguish whether our conditions influence the specificity of the generated feedback, we determined that the participants in the non-spatial Interface left longer feedback that had more stop words. We also uncovered that the conditions produce different categories of feedback. In particular, we found that the spatial Interface tends to generate more *investigation* feedback.

Our findings have several key implications. Evidence from the study shows that granting feedback providers access to history feedback introduces a fixation effect. A fixation effect could mean that a designer who chooses to use a feedback generation tool that allows providers history access will end up with more convergent responses. This fixation effect is especially prominent in non-spatial tools.

We also found that interface classes generate different kinds of feedback. Online feedback system designers in the future should be aware that their decisions influence the generated content. For future work, we would like to include more classes of interfaces and see how their generated feedback differs.

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 [12, 22]. Our study did not consider different stages of design and how they would interact with the choice of design tool. Due to the number of factors we looked at, it was not feasible to include an additional factor of design stage in this study.

The conditions in our study represented two classes of feedback interface without overlap, but there are interfaces that do not match either condition that we studied. For example, in the Voyant and CrowdCrit systems, in addition to offering a free-form response and a text box, the provider could also annotate a region of the design to associate with that comment [3, 9]. 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 whereas novices may value access to history feedback for inspiration [9, 20]. Experts tend to both generate more ideas and to fixate more often [16]. Future work can study any interactions between expertise and interface.

**Limitations**

Our study compared different features of the feedback generated for two classes of interfaces. However, our study did not measure the objective quality of the feedback itself. What we should have done is hire experts to measure the quality of the feedback. Future work could address this limitation by having independent experts rate the perceived quality of the feedback between conditions.

A second limitation is that we recruited feedback providers from the Mechanical Turk platform. The primary incentive for the participant in the study would be financial gain. Future work could test the generalizability of these findings by using crowds driven by different incentives, such as classroom peers, people recruited from social networks, or in the context of online communities such as Reddit.

**CONCLUSION**

Designers are increasingly turning to the use of online tools to collect feedback on n-progress work. Designers have many tools to choose from for collecting feedback for their work online. The design choices within the tool could have consequences for the scope and quality of the feedback generated. But how does the choice of tool influence the scope and quality of the feedback that is received? Our work made three contributions. First, we showed that access to history feedback introduces a fixation effect. Second, we found that while conditions don’t influence feedback specificity, providers using the non-spatial interface left longer feedback. Third, we discover that spatial interfaces tend to generate more *investigation* feedback. We hope this work contributes to building more extensive interfaces for feedback exchange, can help feedback providers better communicate the evaluation of their design, and help designers improve their work through higher quality feedback.

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. *Red Pen.*

5. *Adobe Acrobat.*

6. *Reddit*. Available from: <https://www.reddit.com/r/design_critiques>.

7. *Dribble.*

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Smith, S.M., Ward, T.B., and Schumacher, J.S., *Constraining effects of examples in a creative generation task.* Memory & Cognition, 1993. **21**(6): p. 837-845.

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. 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.

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

17. 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.

18. 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.

19. 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.

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.