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]. Crowd feedback services offer several advantages over alternatives, and are attractive because of their scalability, availability, and affordability [2, 3]. However, one challenge these services encounter is that crowd workers rarely possess skills in specialized domains such as design [4].

To combat this, scaffolding is often employed to provide support to learners as they attempt a novel task. One scaffolding strategy that helps learners express themselves is the use of different media and modes of expression in the interface [5]. A number of crowd feedback systems have implemented this strategy resulting in enhanced quality of collected feedback [3, 6, 7]. This feedback has been shown to lead to improved designs [2].

However, introducing scaffolding into a system may significantly affect people’s behavior by introducing framing effects [8]. For example, asking a feedback provider to annotate a visual design requires them to visually search the design and focus their attention on specific details [3, 9]. Revealing previous feedback to a feedback provider, while encouraging novel ideas, may also encourage conformity [10, 11]. It is therefore important for researchers to understand the relationship between characteristics of the feedback collection interface and the resulting feedback to develop effective scaffolding in crowd feedback systems.

In this paper, we present four interfaces for soliciting design feedback and study characteristics of the collected feedback in each condition. In our experiment we manipulate input type and presence of history to create our interfaces. Modality conditions reflect real-world feedback collection interfaces such as Reddit and Red Pen [12-14]. Meanwhile, the history condition represents another design decision that must be made during the creation of a feedback collection interface that could have important implications on the resulting feedback [15].

We then recruit human participants to provide feedback on visual designs using one of the interfaces. We investigate how choosing a text or annotation feedback input and hiding or revealing history impacts feedback content including number of discrete elements of design referenced and degree of conceptuality. These relationships are explored across a variety of visual designs (poster, static website, and web interface).

# Related work

## Crowd Feedback Systems

## Creativity/Inspiration/Design Feedback

## “Deixis” (indication, pointing out)

# Methodology

Our study examined how two forms of a feedback Interface (non-spatial or 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.

## 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**. The task was priced at $0.50 per entry to reflect minimum wage ($8.25 per hour).

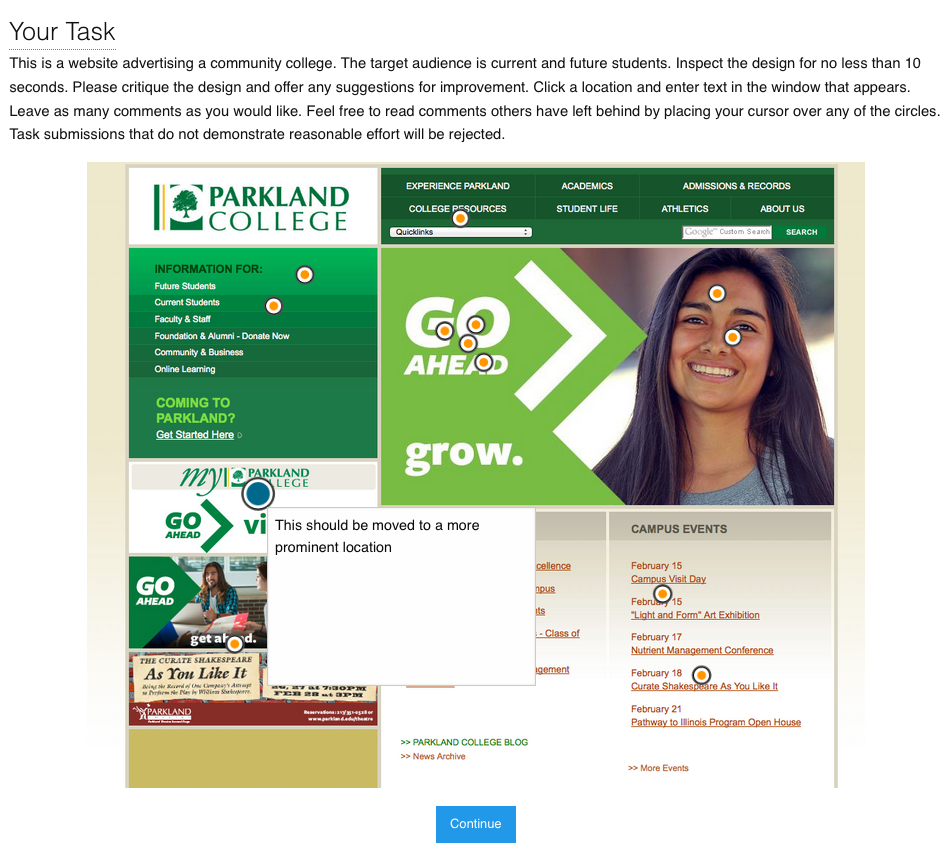
## Designs

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 author of the first design was obtained and the two following designs are public domain.

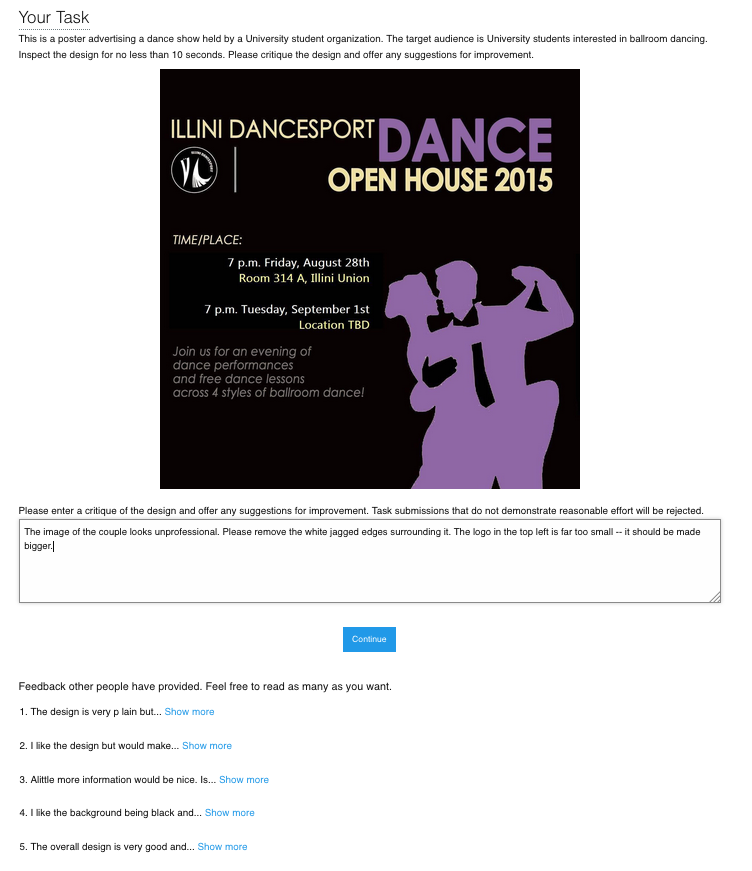
## Feedback Interfaces

The feedback interface features a block of text introducing the task and includes a brief text description of the design and its target audience. The design is then prominently displayed.

In the non-spatial condition, a text area prompting the provider for feedback followed the design image. This text area is accompanied with a button to submit feedback and 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. 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 [12, 14]. We adapted this format however to include a “Show more…” interaction which allowed us to log which pieces of feedback were viewed.



**Figure 1. 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.**

****

**Figure 2. The interface for leaving feedback in the non-spatial condition. A feedback provider enters their feedback in a text area. In the history condition, feedbacks left by previous providers were visible. The participant may choose to view the full feedback by selecting ‘Show more.’**

In the spatial condition, 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. Once committed, a visual marker representing the feedback is 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 the associated visual marker, and could always edit their own feedback by clicking a marker they had laid down. 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. The spatial condition was designed and implemented to reflect popular annotation feedback tools such as Adobe Acrobat and Red Pen [13, 16].

Once satisfied with the feedback, the provider submitted their work.

## Procedure

Upon accepting the task, the feedback provider was presented with a form of consent. If accepted, they were randomly assigned to one of 12 experimental conditions. After leaving feedback in the interface provided, they submitted their work and completed a brief survey.

## Measures

We collected the following measurements.

*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 [4].

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 [17]. 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 similarity between the recent comment and the feedback that was seen and the feedback that was not seen. The Python pattern.en toolkit cosine text distance implementation was used to compute this measurement.

In addition to the above, we also measured general 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, a provider completed a self-assessment survey. The survey had 5 structured responses. They were asked to answer three questions on a five point Likert-scale to rate their design expertise, perceived effort, and perceived usefulness of feedback, with a score of 5 as the most favorable. The survey also included two questions for demographics.

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

**Content Analysis**

*Non-spatial feedback had more stop words*

ANOVA detected no effect on any of the conditions 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).

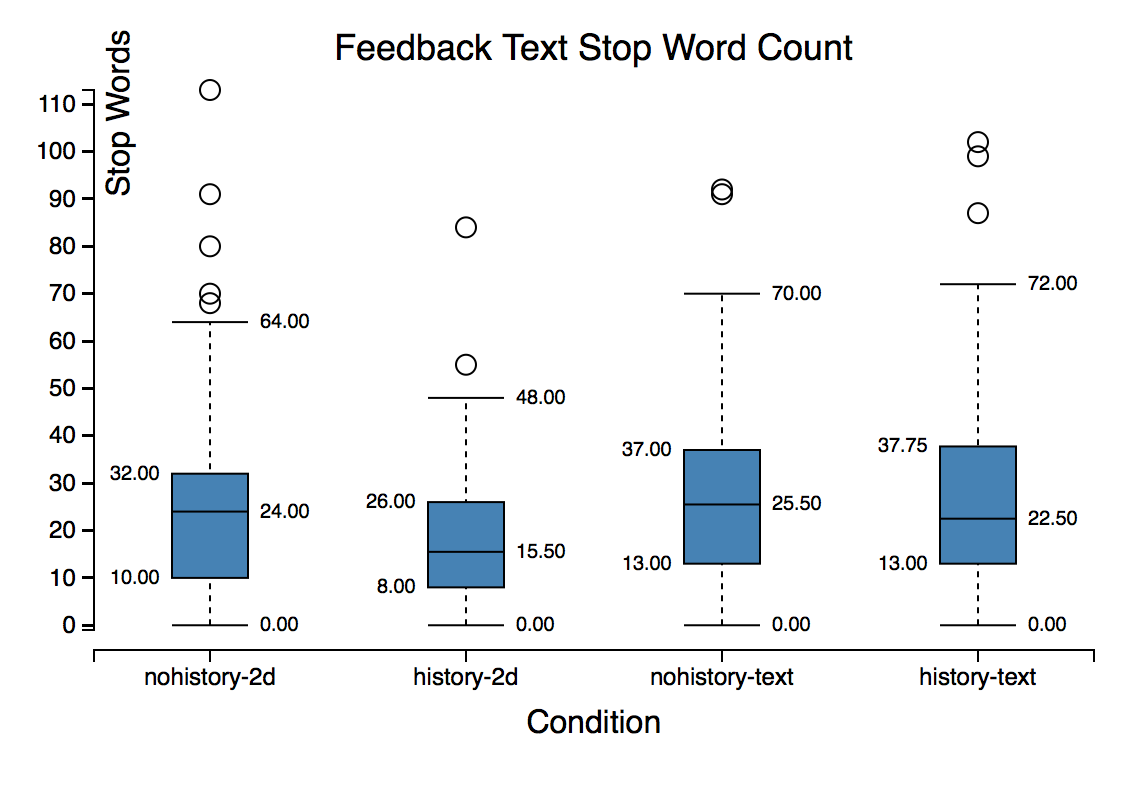
Digging deeper, ANOVA uncovered a main effect of Interface on stop word count (F(3,357)=6.93; p=0.0089). Figure 1 summarizes stop word count. Tukey’s HSD showed that stop word count in the non-spatial condition (μ=27.31) was greater than the spatial condition (μ=21.98; p=0.0084).

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.”*

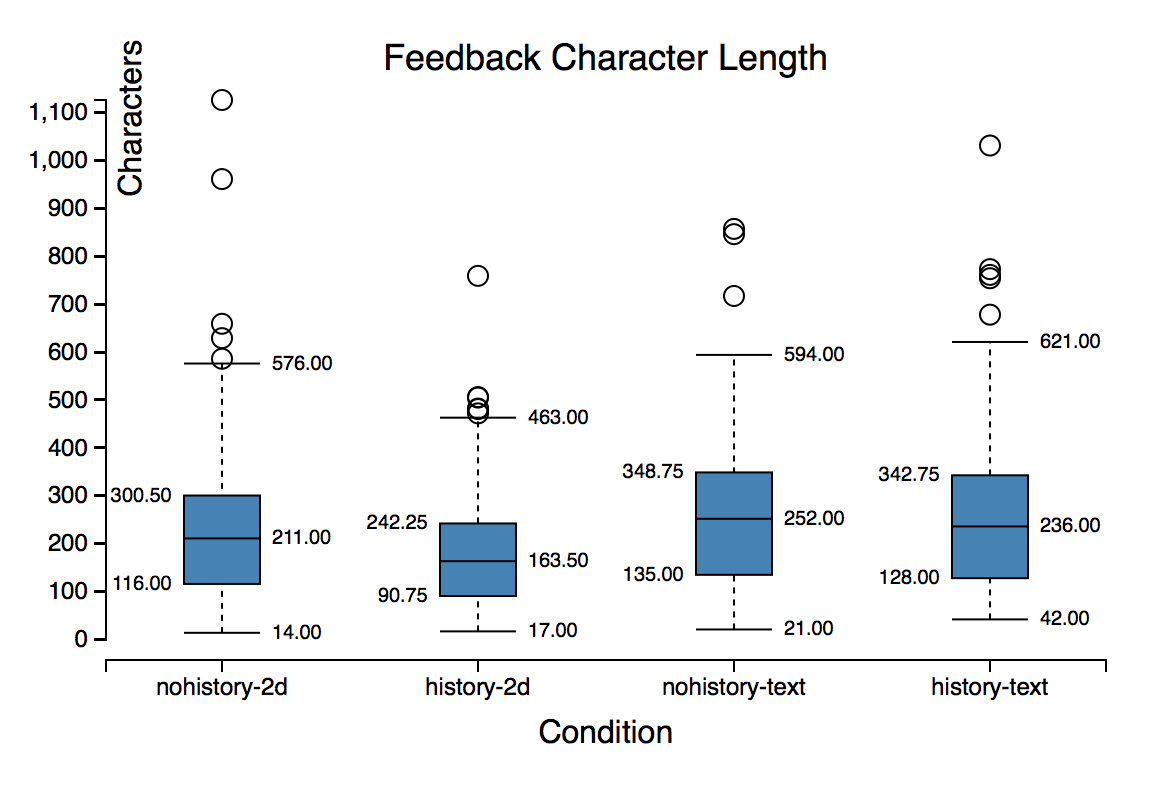
*Non-spatial condition produced longer feedback*

An ANOVA revealed that Interface had a main effect on character length (F(3,357)=7.86; P=0.0053). Character length 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=269.65 characters) was longer than the feedback from the spatial condition (μ=217.37; p=0.0051). No other effects were discovered.

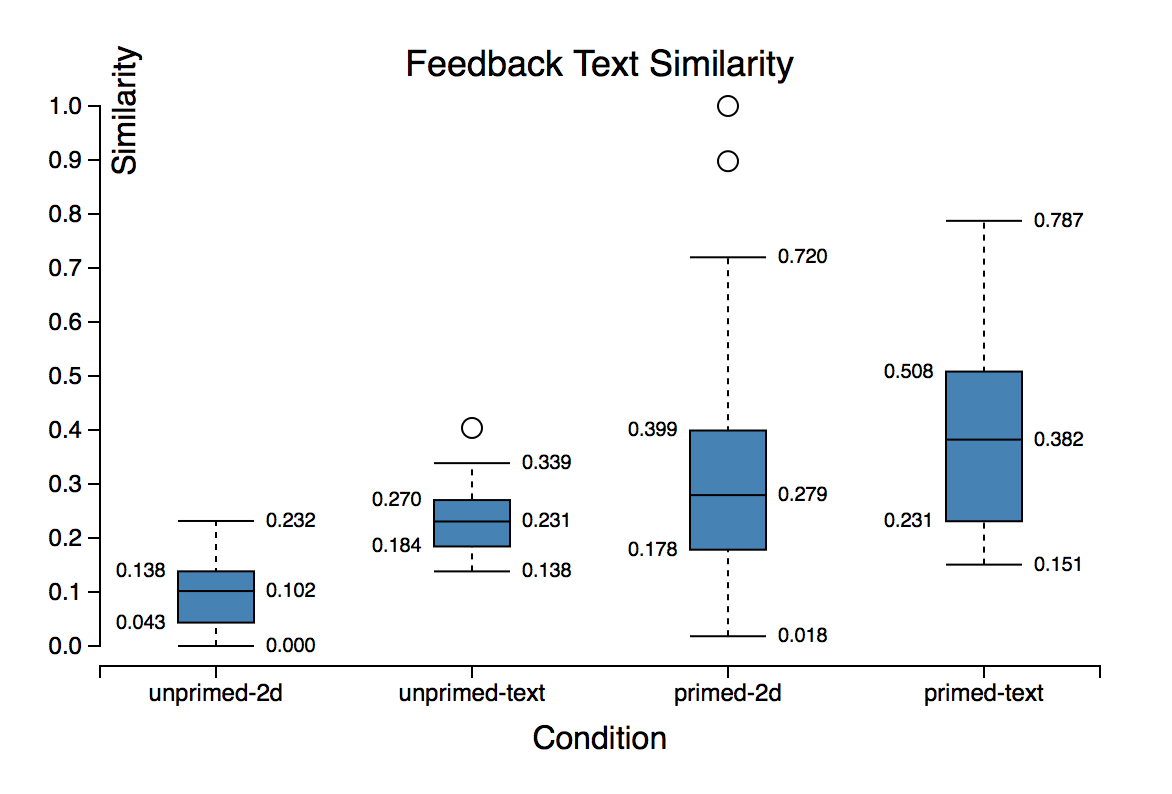
The non-spatial condition may have led to longer feedbacks due to the need for use of deixis, i.e. words or phrases such as “here” or “there” that require further contextual information to be understood but eliminate the need for explicit description of the visual elements referenced by feedback.



**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.**



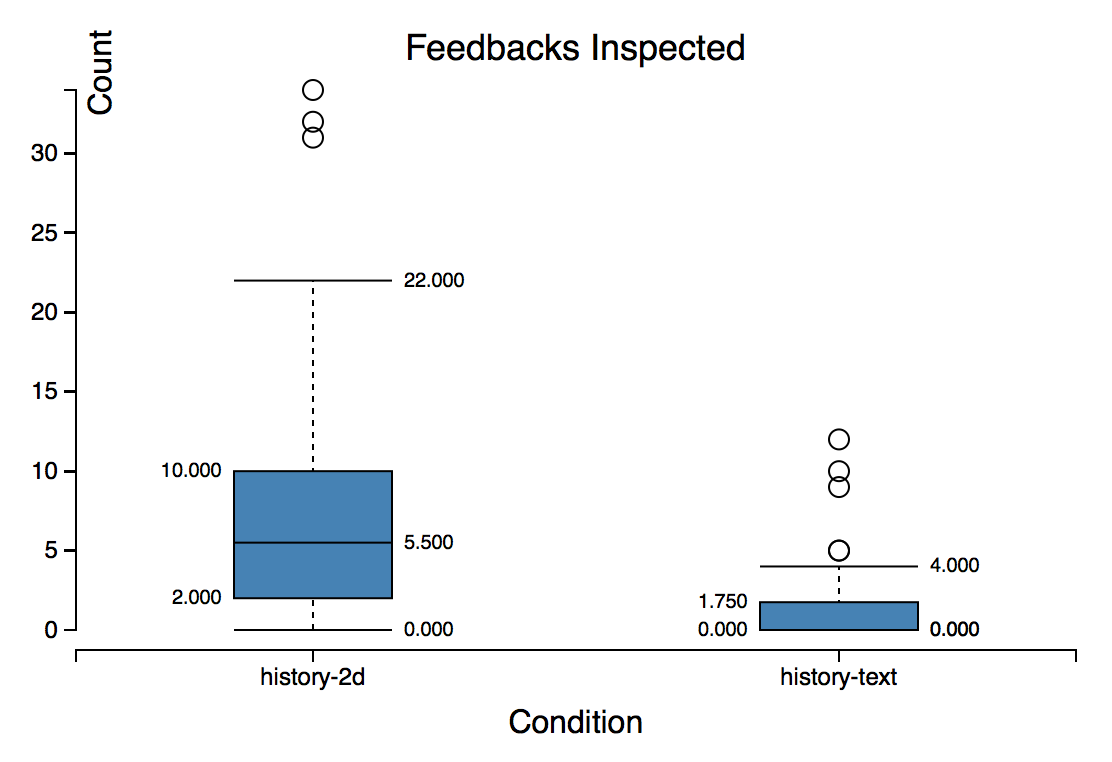
**Figure 3. Similarity ranking of generated feedbacks compared to viewed and unviewed history by condition. Generated feedback was more similar to viewed history.**

**Behavioral Measures**

*Providers inspect more feedback in spatial condition*

The number of instances of feedback inspected by providers is visualized in Figure 4. ANOVA revealed a main effect of Interface on instances of feedback inspected (F(3, 180)=60.57; p=0.0001). Digging further, Tukey’s HSD showed that spatial condition providers inspected more feedback instances (μ=7.29) than the non-spatial condition (μ=1.14, p=0.0001).

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

**

**Figure 4. 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.**

*Generated feedback was more similar to viewed history*

For calculating similarity of generated feedback to history, we only considered instances of generated feedbacks where the provider had examined some prior history beforehand. This left us with 200 instances of feedback in the spatial condition and 42 instances of feedback in the non-spatial condition. Figure 3 displays feedback similarity scores.

ANOVA revealed a main effect indicating that generated feedbacks were more similar to those that providers inspected than other present feedbacks that weren’t inspected by the provider (F(3,230)=26.59; p=0.0001). Tukey’s HSD shows that similarity between generated feedback and inspected feedback (μ=0.11) was higher than generated feedback and feedback that wasn’t inspected (μ=0.044; p=0.0001).

This suggests that presence of a history introduces a fixation effect for the feedback providers. This effect is analogous to the fixation effect to pictorial representations and design examples [citation needed].

*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 condition (μ=0.069; p=0.0039). This effect is also visible in Figure 3.

This may be due to the fleeting nature of the spatial condition history interface, which requires the user to be hovering a visual marker to unveil history as opposed to the text condition where the history is visible until one opts to once again hide the history feedback.

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).

**Self-Assessment**

*Design influenced perceived feedback usefulness*

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 university student, whereas Designs B and C were professional web pages representing professional entities.

ANOVA detected no effect of any of the conditions on self-assessment of provider effort (μ=3.2; σ=1.1).

## Discussion

# Future work

# Conclusion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Condition** | | | |
| 0 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 3 |  |  |  |  |
| 2 |  |  |  |  |
| 1 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 6 |  |  |  |  |

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. 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.

5. Soloway, E., M. Guzdial, and K.E. Hay, *Learner-centered design: the challenge for HCI in the 21st century.* interactions, 1994. **1**(2): p. 36-48.

6. 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.

7. 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.

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. 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.

10. 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.

11. 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.

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

13. *Red Pen.*

14. *Dribble.*

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. *Adobe Acrobat.*

17. 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).