Comparing Text and Visual Annotation Survey Tools for Feedback Collection

| 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 [3]. Crowd feedback services offer several advantages over alternatives, and are attractive because of their scalability, availability, and affordability [1, 2, 4]. However, one challenge these services encounter is that crowd workers rarely possess skills in specialized domains such as design [5].

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 [6]. A number of crowd feedback systems have implemented this strategy resulting in enhanced quality of collected feedback [4, 7, 8]. This feedback has been shown to lead to improved designs [1].

However, introducing scaffolding into a system may significantly affect people’s behavior by introducing framing effects [11]. 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 [9, 10]. Revealing previous feedback to a feedback provider, while encouraging novel ideas, may also encourage conformity [13, 14]. 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 [15, 16, 17]. 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 [12].

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 input modality (text or annotation based) and history (present or absent) affect feedback content. We address the following research questions:

* RQ1: How do modality and access to history influence the generated feedback content?
* RQ2: How do modality and access to history influence performance metrics and provider self-assessment?
* RQ3: Does history encourage convergent or divergent feedback?

These questions reflect choices that real world feedback collection system implementers must make but that nobody has compared. Our work will create the empirical body of knowledge about how these choices affect properties of the collected feedback.

## Experimental Design

To answer these questions, we conducted a full-factorial, between-subjects experiment. The factors were Modality (Text vs. Annotation) x History (Present vs. Absent) x Design (Poster vs. Static Website vs. Web Application), giving a 2x2x3 design.

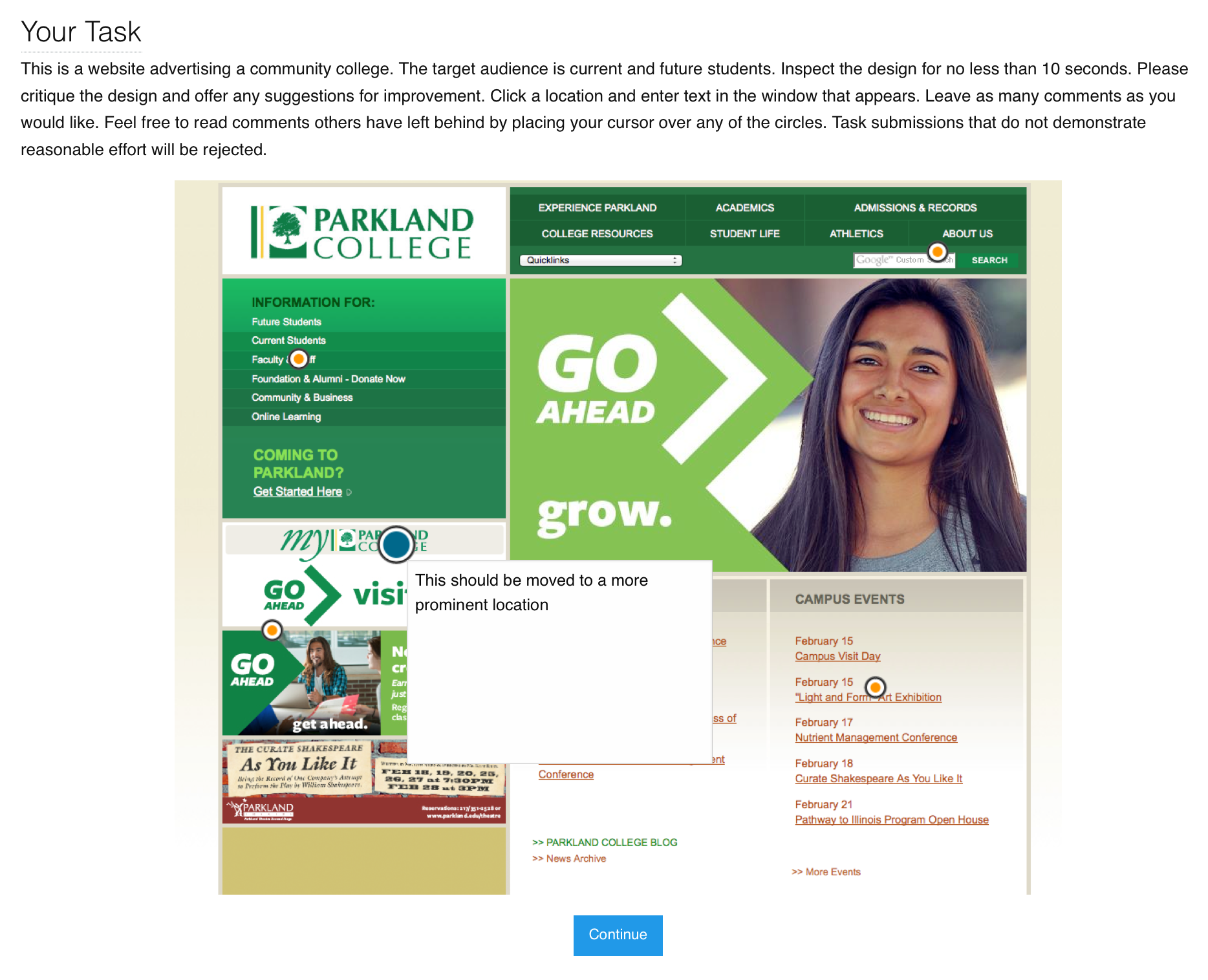
## Participants

Feedback providers (N=360) were recruited from Mechanical Turk. Insert demographic data here. The task was priced at $0.50 per entry to reflect Illinois minimum wage ($8.25 per hour).

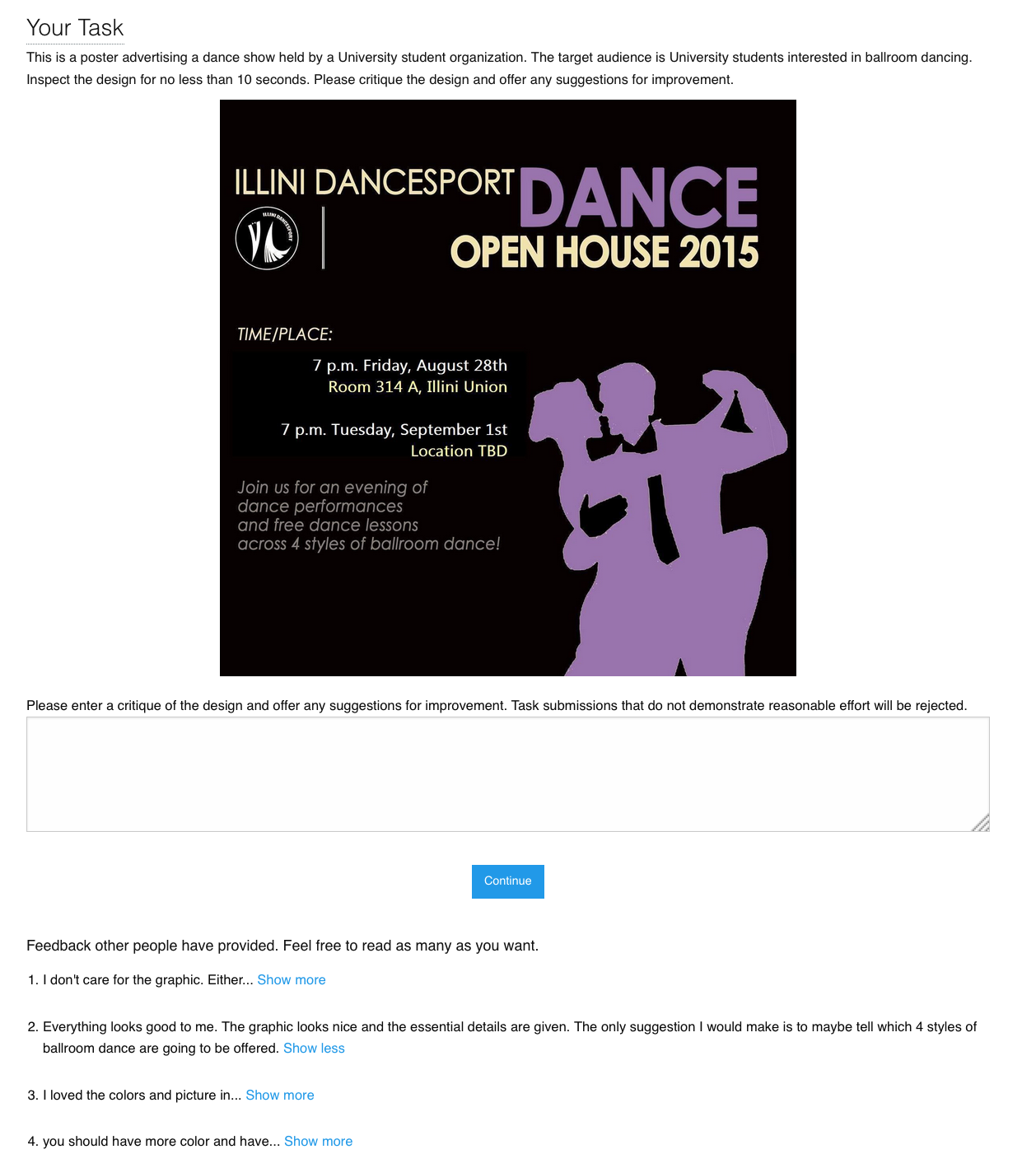
## Designs

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.

While these elements of the interface are present for all conditions, the interfaces for specific conditions are modified to manipulate input modality and presence of history.



**Figure 1. The interface for leaving feedback in the annotation condition. Providers can leave comments by selecting a region on the design and entering text in a window. They could leave as many comments as they wanted and were allowed to look at the history by hovering existing markers.**



**Figure 2. The interface for leaving feedback in the text condition. Feedback providers may use a text area to input feedback text. In the history condition, feedbacks left by previous providers were visible. The participant may choose to view the full feedback by selecting ‘Show more.’**

## Task 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 text 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. When crossed with the history condition, past feedback was displayed underneath this form. The first few words of a past feedback were displayed and accompanied by a “Show more...” link inviting the provider to reveal more of any feedback that caught their attention. The rationale behind this design was twofold – the survey represented a common feedback collection paradigm that affords text input such as Reddit or Dribble [15, 17], and the “Show more…” interaction provided us with access to the history of feedbacks the provider selected.

The annotation condition prompts providers to leave feedback by annotating the presented design. To leave feedback, a provider selected a location on the design image. The system then displayed a small window into which the provider entered text. Entering feedback leaves a visual marker overlaying the design. The provider was allowed to hover over any visual markers to reveal the annotated feedback and could edit their own feedbacks by clicking the marker. Following the design, a button to submit feedback and complete the task was displayed. When crossed with the history condition, past feedbacks were displayed as visual markers overlaying the design. The provider would be free to hover over any of the visual markers to reveal previously left feedback but could not edit them. To leave feedback, the provider could click a location on the design to leave an annotation just like in the no-history condition. The annotation modality condition was designed and implemented to reflect popular annotation feedback tools such as Adobe Acrobat and Red Pen [16, 17].

## Procedure

Upon accepting the task, feedback providers were assigned in order to one of the 12 conditions. Each condition was displayed as previously described in the experiment interface section. The provider was instructed to critique the design and offer any suggestions for improvement. The text introducing the task was manipulated based on condition to best reflect instructions for the interface to which the worker was assigned.

## Measures

We collected the following measurements.

*Content analysis*

Feedback content was categorized for each condition. Responses were partitioned into idea units and these idea units were further classified based on system used previously by researchers [18]. Two coders, selected for their experience in HCI, then assigned the labels. Cohen’s Kappa, a measure of reliability between multiple raters and categories, was 0.81 on training samples. Coders were paid $25 for their effort.

A content specificity metric was also calculated for each feedback provided. Content specificity was measured 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”). The specificity metric was normalized to range from 0.0 to 1.0. In the past, other researchers have used this technique [5].

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

*Content similarity*

In the history conditions, all interactions between the feedback provider and feedbacks left by previous providers were recorded. For the annotation condition, we logged each time the provider revealed a previous feedback by hovering over the visual markers. In the text condition, we logged each time the provider selected the “Show more…” link.

Next, for each feedback provider, we aggregated the prior feedback that they viewed. Each time they left feedback, we compared the similarity of the provided feedback with those that they had and had not seen.

The Python NLTK toolkit was employed to calculate distance between feedback texts to compute content similarity for each participant individually.

*Self-assessment*

Following the design critique task, providers were prompted to complete a self-assessment. Effort exerted was collected on a five point Likert-scale ranging from low effort (1) to high effort (5). Similarly, the perceived usefulness of their feedback was assessed on a five point Likert-scale ranging from not useful (1) to very useful (5).

*Behavioral measures*

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.

# Results

In total, 30 submissions were collected per condition for a total of 360 submissions. Due to high variability, responses that were two standard deviations for away from mean for each measure were excluded.

**Content Analysis**

*Text input produced more specific feedback*

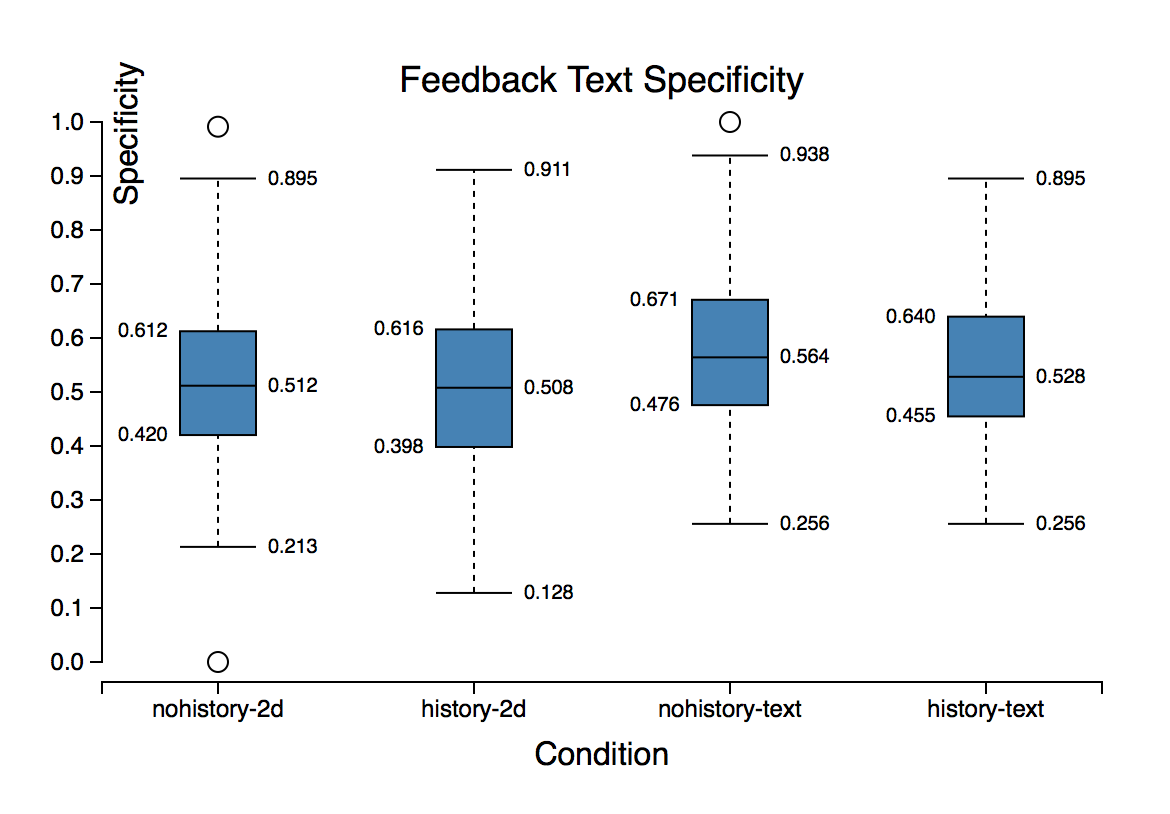
Graph 1 summarizes feedback specificity. In calculating specificity, feedback for which the Wordnet structure was insufficient were thrown out, leaving us with 25 entries of specificity per condition. Performing ANOVA revealed a main effect of modality on feedback text specificity (F(3,300)=7.7678; p=0.0057). Tukey’s HSD unveiled that the specificity of the annotation condition (μ=0.5498) was lower than that of the text condition (μ=0.6044; p=0.0053).

One possible explanation for this is that the additional spatial context information provided by the visual condition reduces the need for the specific language necessary to convey the same information in the text condition.

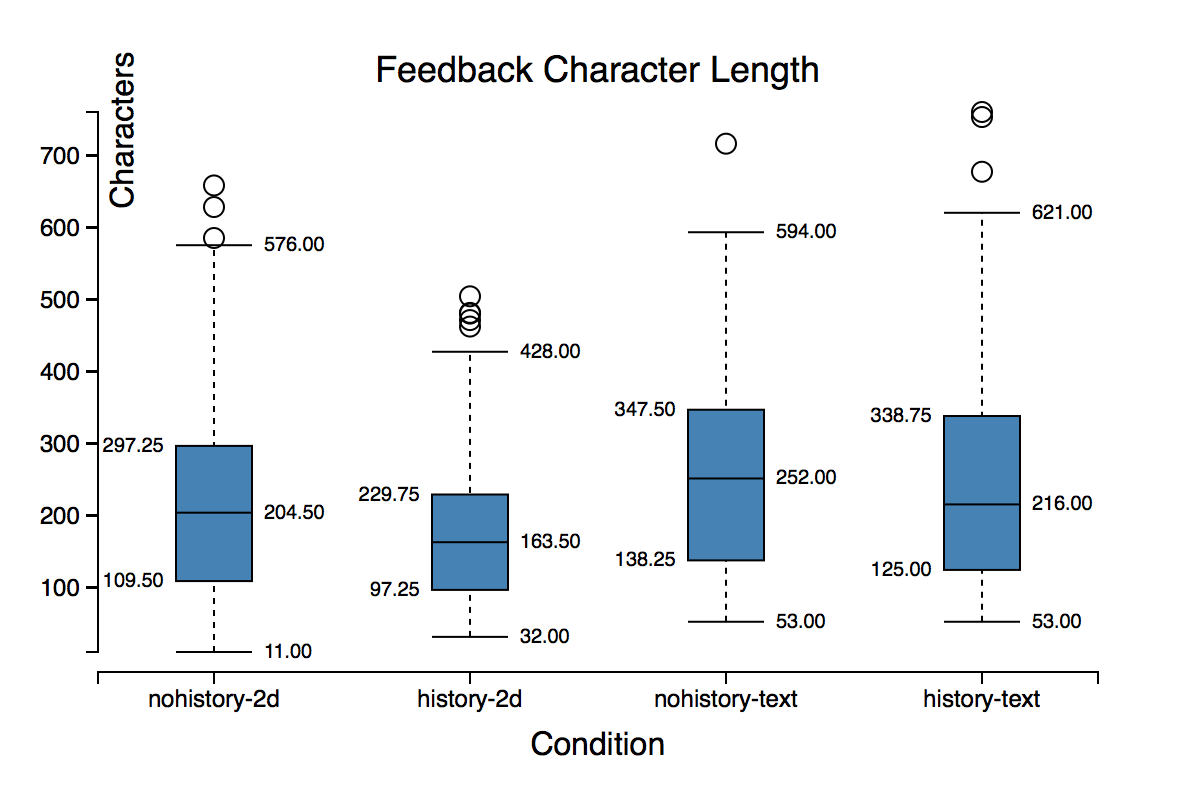
*Text input produced longer feedback*

An ANOVA revealed that modality had a main effect on character length (F(3,336)=12.6333; P=0.0004). Character length per condition can be seen in Graph 2. Pairwise comparison using Tukey’s HSD showed that the length of the feedback in the visual condition (μ=201.1 characters) was shorter than the feedback from the text condition (μ=256.5; p=0.0005). No other effects were discovered.

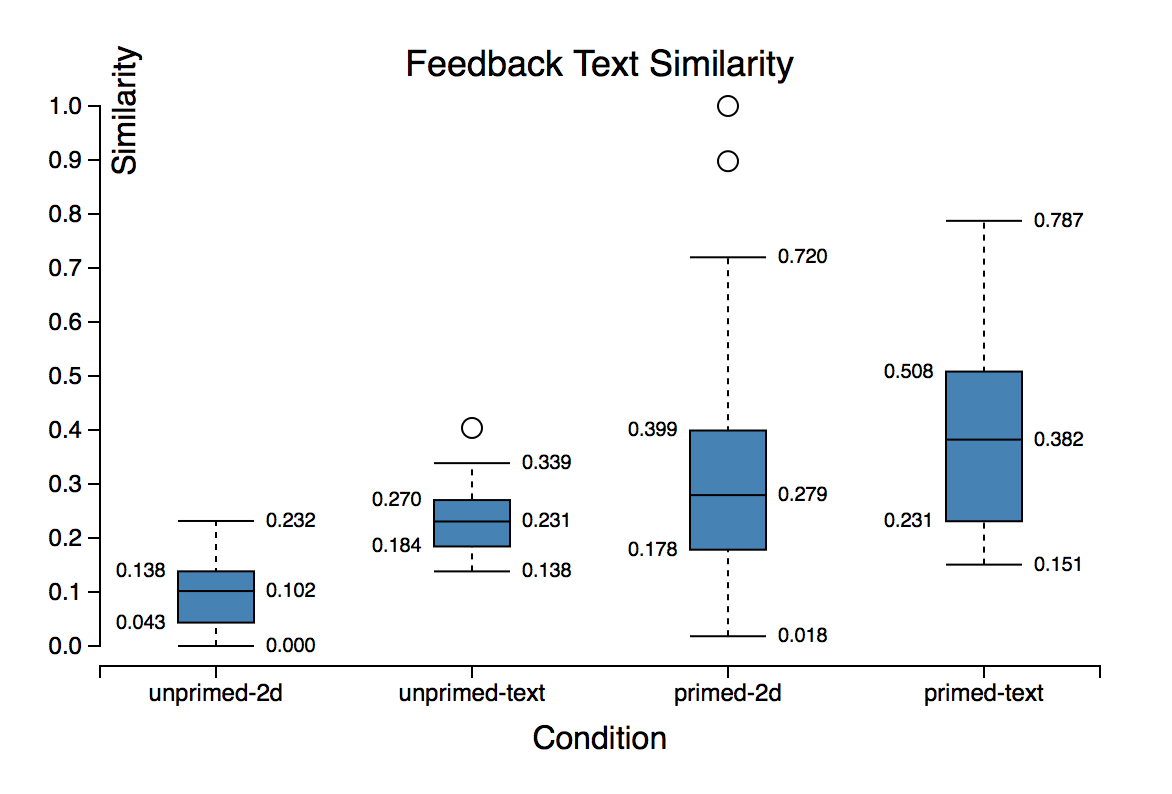
The visual condition may have led to shorter feedbacks due to the use of deixis, i.e. words or phrases such as “here” or “there” that require further contextual information to be understood but eliminate the need for explicit description of the visual elements referenced by feedback.



**Graph 1. Text specificity ranking of feedback content by condition. Text modality is more specific.**



**Graph 2. Cumulative character length of feedback content by condition. Text modality has longer feedbacks.**



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

**Content Similarity**

*Generated feedback was more similar to viewed history*

Graph 3 displays feedback similarity rankings. For calculating similarity, only instances in which feedback providers spent time inspecting other feedbacks were considered, leaving us with 60 entries total. Three way 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(3,60)=12.6565; p=0.0009). Tukey’s HSD shows that similarity between generated feedback and inspected feedback (μ=0.2939) was higher than generated feedback and feedback that wasn’t inspected (μ=0.1620; p=0.0022).

This would suggest that the history condition encourages convergent thinking, as previous feedback primes participants to write about similar features.

*Text condition feedback was more similar to viewed history*

An additional main effect revealed by three way ANOVA was the influence of modality on similarity to viewed history (F(3,60)=12.3230; p=0.0010). Tukey’s HSD showed that similarity to viewed feedbacks in the text condition (μ=0.29) was higher than that of the visual condition (μ=0.16; p=0.0025). This effect is also visible in Graph 3.

This may be due to the fleeting nature of the visual 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.

**Self-Assessment**

*Conditions had no effect on self-assessment*

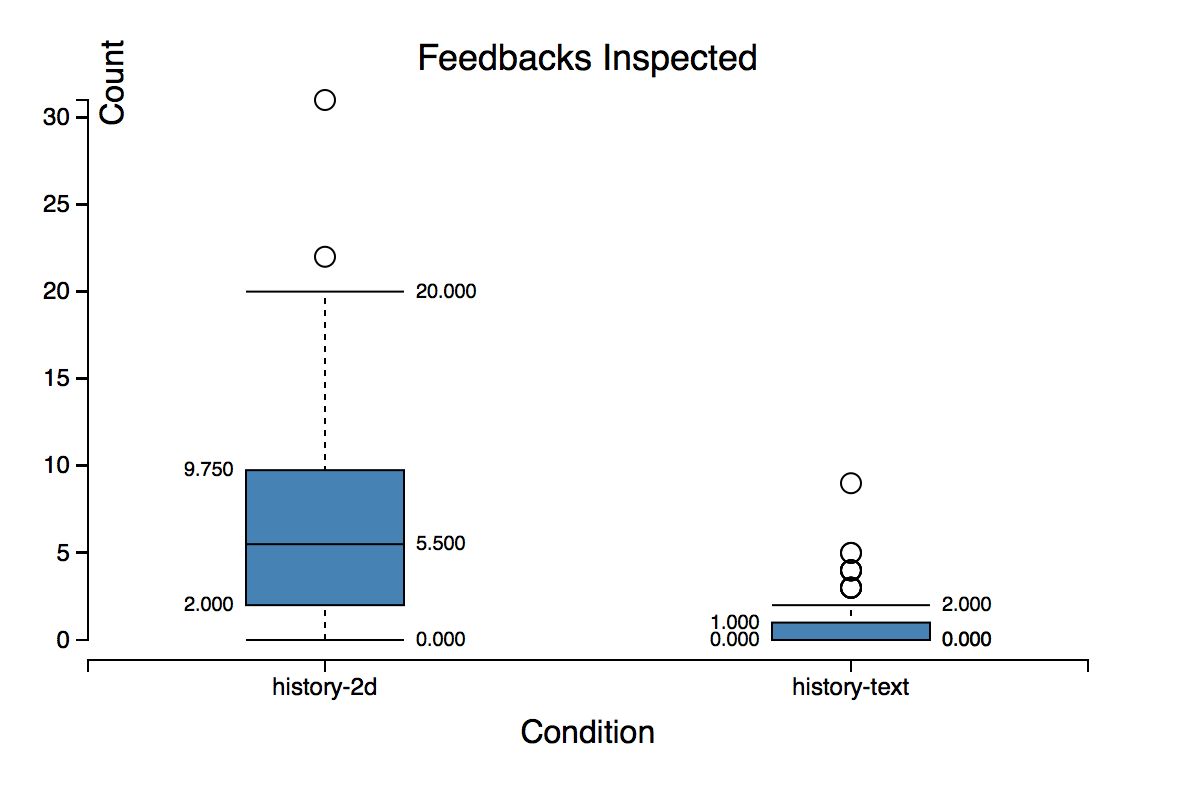
Three way ANOVA detected no effect of any of the conditions on self-assessment of provider effort and perceived usefulness.

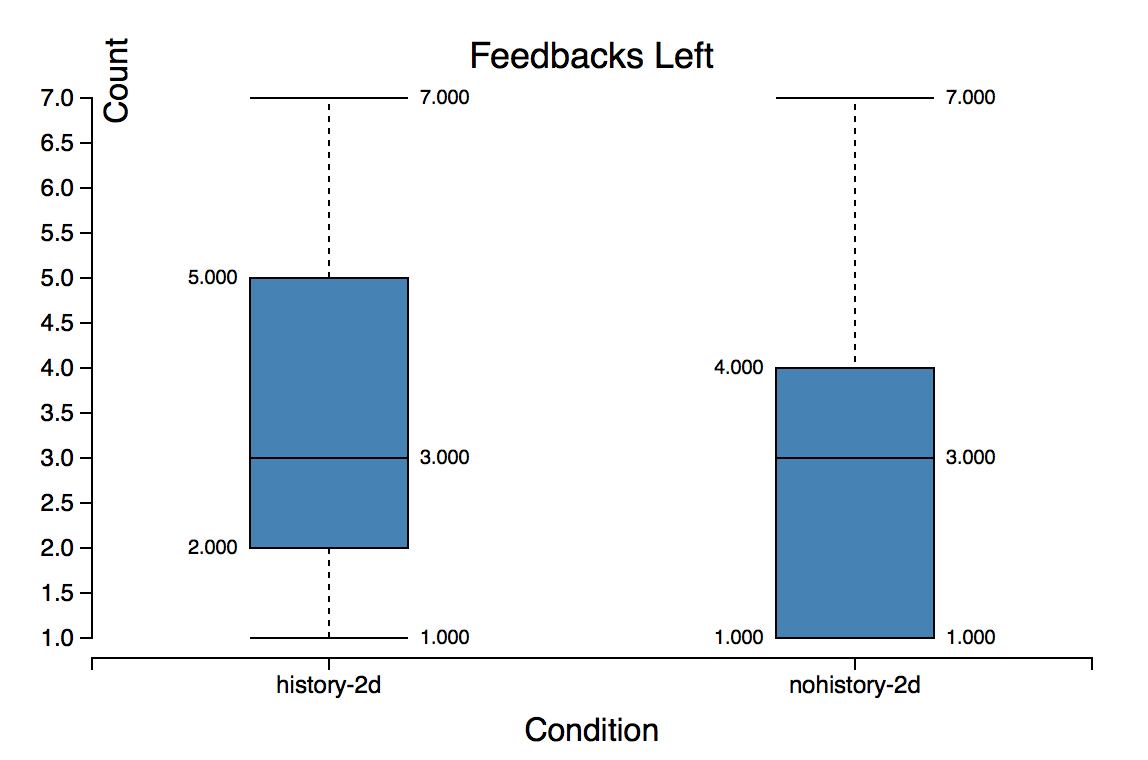
**Behavioral Measures**

*Providers inspect more feedback in visual condition*

The number of instances of feedback inspected by providers is visualized in Graph 4. Three way ANOVA revealed a main effect of modality on instances of feedback inspected (F(3, 168)=87.2398; p=0.0001). Digging further, Tukey’s HSD showed that visual condition providers inspected more feedback instances (μ=6.65) than the text condition (μ=0.90, p=0.0001).

One explanation for this effect is the ease with which one can inspect history feedback in the visual condition relative to the text condition. Whereas providers must click the ‘Show more…’ link to fully reveal history feedback in the text condition, the visual condition merely required the provider to hover the visual marker associated with the feedback instance to unveil it.

 **Graph 4. History conditions inspected by modality. Providers inspected more feedback instances under the visual condition.**



**Graph 5. Visual condition feedback generated by history condition. Providers generated more instances of feedback when exposed to history.**

*History leads to more feedback instances under visual condition*

Participants had the opportunity to leave multiple instances of feedback in the visual condition. Graph 5 renders the count of feedback instances generated under history and no-history conditions. Two way ANOVA with design and modality as factors revealed a main effect of modality on count of instances of feedback generated (F(2, 168)=4.0167; p=0.0467). Tukey’s HSD showed that providers in the history condition left more feedback instances on average (μ=3.36) than those in the no-history condition (μ=2.85, p=0.0465).

It is possible that the ability to see history feedback served as motivation for the providers to leave behind more instances of their own feedback.

*Conditions had no effect on task completion time*

Three way ANOVA revealed no significant effect of the conditions on task completion time.

## Discussion

# Future work

# Conclusion

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

# REFERENCES

1. Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. 2015. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15). ACM, New York, NY, USA, 1637-1648.
2. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). ACM, New York, NY, USA, 1433-1444.
3. XZZZZ Elkins, J. *Art Critiques: A Guide.* New Academia Publishing, Washington DC, 2012.
4. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15). ACM, New York, NY, USA, 473-485.
5. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Vennix, Steven P. Dow, Björn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques . To appear in Proceedings of CSCW 2016 .
6. XZZZZZ Elliot Soloway, Mark Guzdial, and Kenneth E. Hay. 1994. Learner-centered design: the challenge for HCI in the 21st century. interactions 1, 2 (April 1994), 36-48.
7. Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15). ACM, New York, NY, USA, 235-244.
8. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). ACM, New York, NY, USA, 1433-1444.
9. Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). ACM, New York, NY, USA, 227-236.
10. William C. Hill and James D. Hollan. 1991. Deixis and the future of visualization excellence. In Proceedings of the 2nd conference on Visualization '91 (VIS '91), Gregory M. Nielson and Larry Rosenblum (Eds.). IEEE Computer Society Press, Los Alamitos, CA, USA, 314-320.
11. Catherine M. Hicks, Vineet Pandey, C. Allie Fraser, Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment. (<http://pages.ucsd.edu/~cmhicks/Projects/pdf/feedback.pdf>)
12. Kulkarni, C., Dow, S., and Klemmer, S. R. 2012. Early and repeated exposure to examples improves creative work. In Proceedings of Cognitive Science
13. Buxton, B. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann, 2010
14. XZZZZ Smith, S.M., Ward, T.B., and Schumacher, J.S. Co n- straining effects of examples in a creative generation task. Memory & Cognition 21, 6 (1993), 837 – 845
15. Reddit. <https://www.reddit.com/r/design_critiques>
16. Red Pen. <https://redpen.io/>
17. Dribble. <https://dribble.com>
18. XZZZZZZ Deanna P. Dannels and Kelly Norris Martin. 2008. Critiquing Critiques: A Genre Analysis of Feedback Across Novice to Expert Design Studios. Journal of Business and Technical Communication, 22, 2: 135- 159. http://dx.doi.org/10.1177/1050651907311923