DESIGNING WITH CROWDS

BY

ANBANG XU

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Brian P. Bailey, Chair
Associate Professor Wai-Tat Fu
Associate Professor Karrie Karahalios
Assistant Professor Steven P. Dow, Carnegie Melon University
Abstract

Feedback plays an essential role in the evolution of a design from initial concept to final representation. There are several approaches to receive feedback on designs such as organizing critiques, informally asking peers, and participating in online feedback communities. However, these social approaches highly depend on a user’s social capital and peer networks, and thus the access to desired feedback is limited.

This dissertation explores the use of non-expert crowds to generate helpful, affordable, and timely feedback for visual design. The key innovation is the concept of generating structured feedback from non-experts. Based on a formative study, I identified several types of perception-oriented feedback desired from non-experts. To generate this feedback, I designed and implemented a new crowd-based system called Voyant. Voyant enables any user who creates visual designs to receive perception-oriented feedback on the designs from a selected audience. The generated feedback includes the elements seen in a design, the order in which elements are noticed, impressions formed when the design is first viewed, and interpretation of the design relative to guidelines in the domain and the user’s stated goals. I evaluated Voyant with users and their own designs. My evaluations reveal that crowd feedback helps users improve their designs in practice. The feedback helped users identify design problems, gain insight into their designs, and consider their designs from diverse perspectives. The monetary cost of generating the feedback was considered a fair tradeoff for not having to organize critiques or interrupt peers. The content analysis reveals that Voyant generated more interpretative, diverse, and critical feedback than free-form prompts.

This dissertation provides the empirical evidence that crowd feedback benefits the end user and contributes new lessons for designing crowd feedback systems that support design activities. Crowd feedback systems provide users more opportunities to make design decisions based on audience input rather than solely on intuitions. These systems have the potential to tighten the design cycle and help users iterate toward solutions that better connect with the intended audience.
to my parents

who taught me to love life, people, and science
Acknowledgments

I would like to thank my advisor, Prof. Brian P. Bailey, for the patient guidance and invaluable support over the past six years. You helped tremendously to shape the thinking that went into this dissertation. This dissertation would not have been possible without your advice, criticism, and generous support. I am truly grateful for the time you spent with me in countless consultations, and feedback on publications.

I would to express my gratitude to Wai-Tat Fu, Karrie Karahalios, and Steven Dow for valuable advice of my research. Whenever needed, you have happily provided ideas and suggestions. Your suggestions are very helpful to me long into my research career. I am deeply thankful.

My thanks go to Tara Matthews and Jacob Biehl, who gave me great opportunities to conduct research in industry. You are constantly available for any question I had over the past a few years. I also would like to thank Jilin Chen, Michael Muller, Hernan Badenes, Barton A. Smith, Michelle X. Zhou, Eleanor Rieffel, Thea Turner, William Melle, Huaming Qu, Ping Guo, and Li Hua. It has been an honor to work with these wonderful mentors. Your strong support and suggestions have given me the best possible start to my research career.

I would like to also acknowledge Joshua Hailpern, Moushumi Sharmin, Sanny Lin, Roshanak Zilouchian, Nikita Spirin, Dae Hoon Park, Brett Jones, Raj Sodhi, Scarlett Herring, Brittany Smith, Andrew Harris, Vera Liao, Tony Bergstrom, Eric Gilbert, Shamsi Iqbal, Arjun Venkataswamy, Shih-Wen Huang, Huaming Rao, Xiao Yang, John Lee, Wayne Wu and Grace Yen. Thank you for aiding me in my research. Your friendship has been invaluable.

My parents have done so much for me. You are the best parents anyone could have asked for. Thank you for teaching me how to go after my dreams. From first grade to my Ph.D., you were always there. Without you I would not be where I am today. I am grateful to you every day.

Special thanks to my wife. You are the love of my life, my soul mate and my very best friend. It is beautiful to grow together. From the bottom of my heart, thank you for the endless love, support, and encouragement you have given me.
# Table of Contents

List of Tables ........................................................................................................................................... ix
List of Figures ........................................................................................................................................... x

Chapter 1 Introduction ................................................................................................................................. 1
  1.1 An Example of Why Audience Feedback Is Needed ................................................................. 1
  1.2 Existing Options for Receiving Feedback .................................................................................. 2
  1.3 A Crowd-based System for Generating Feedback ................................................................. 3
  1.4 Scope ........................................................................................................................................... 4
  1.5 Contributions ................................................................................................................................. 4

Chapter 2 Literature Review ....................................................................................................................... 7
  2.1 Crowdsourcing in Design ........................................................................................................... 7
  2.2 The Creative Need for Design Feedback ..................................................................................... 10
  2.3 Theory of Feedback in Design Education .................................................................................. 11
  2.4 Methods for Receiving Design Feedback ................................................................................ 14
  2.5 Representation of Design Feedback .......................................................................................... 16
  2.6 Studies of Q&A and Online Creative Communities .................................................................. 18

Chapter 3 Preliminary Work ....................................................................................................................... 20
  3.1 Analysis of a Large Online Critique Community .................................................................... 20
  3.1.1 Feedback Community Studied ............................................................................................... 22
  3.1.2 Research Questions and Methodology ................................................................................. 23
  3.1.3 Benefits of Online Feedback ................................................................................................. 25
  3.1.4 Users’ Expectations of Online Feedback .............................................................................. 28
  3.1.5 Feedback Reciprocity ........................................................................................................... 36
  3.1.6 Design Implications for Online Feedback Communities ................................................... 40
  3.2 Crowdsourcing Design Feedback ............................................................................................... 42
  3.2.1 Sequential Process for Feedback ......................................................................................... 43
  3.2.2 Crowdsourcing Model for Online Feedback ......................................................................... 44
  3.2.3 Testing My Approach ............................................................................................................ 46
  3.3 What Was Learned ....................................................................................................................... 48
8.2.1 Research Questions ................................................................. 99
8.2.2 Methodology ........................................................................... 99
8.2.3 Changes on Design Artifact .................................................... 100
8.2.4 Validity of Crowd Feedback ..................................................... 108
8.3 Content Analysis ......................................................................... 111
  8.3.1 Research Questions ............................................................... 111
  8.3.2 Data Collection ...................................................................... 113
  8.3.3 Feedback Genres .................................................................. 113
  8.3.4 Topic Diversity ...................................................................... 116
  8.3.5 Linguistic Styles .................................................................... 118
8.4 Summary ..................................................................................... 123

Chapter 9 Discussion and Future Work .............................................. 124
  9.1 Finding Problems ...................................................................... 124
  9.2 Improving Feedback Quality ..................................................... 125
  9.3 Measuring Feedback Reliability ............................................... 125
  9.4 Optimizing Time and Cost ........................................................ 127
  9.5 Supporting Comparison of Feedback on Multiple Designs .......... 128
  9.6 Reflections on Use Cases and Users ......................................... 128
  9.7 Other Feedback Types and Crowds .......................................... 129
  9.8 Other Studies of Free-form Feedback and Structured Feedback .. 130
  9.9 Crowdsourcing Feedback in Other Creative Domains ............... 130
  9.10 Hybrid Combination of Crowd and Machine ......................... 131
  9.11 Example Search based on Crowd Feedback ............................ 132
  9.12 Managing Alternatives on a Large Scale ................................. 133
  9.13 Social and Cultural Factors in Crowd Feedback Systems .......... 134

Chapter 10 Conclusion ...................................................................... 136

Appendix A: Consent Form ............................................................... 138
Appendix B: Formative Study Interview ........................................... 142
Appendix C: Lab Study Interview ..................................................... 145
Appendix D: Lab Study Survey ......................................................... 147
Appendix E: Classroom Study Survey .............................................. 149
List of Tables

Table 3.1. Descriptives of user activity in our data set.......................................................... 24

Table 3.2. The main themes and a sample of the questions asked within each theme for the semi-structured interviews.................................................................................................................. 25

Table 3.3. Stepwise regression was applied to test how different variables of participation predict the change of a user’s photo rating. Four of the six variables tested were significant. The coefficients are from the final model while the R$^2$ values are from each regression step. The model included the initial photo rating which is inversely associated with the change in photo rating. This may reflect a regression-to-the-mean effect................................................................. 28

Table 4.1 The main topics and a sample of questions asked within each topic for the semi-structured interviews.................................................................................................................. 50

Table 6.1. The typical number of workers recruited and cost for each workflow in the current implementation......................................................................................................................... 74

Table 7.1. Implementation Statistics of Voyant............................................................................ 81

Table 8.1. Participant ratings for the different types of feedback generated and the overall system................................................................................................................................. 89

Table 8.2. The distribution of the four types of changes made by participants according to crowd feedback...................................................................................................................... 99

Table 8.3. Free-form feedback and Voyant feedback generated on a design......................... 112

Table 8.4. A sample of responses from free-form and structured feedback (n = 1000 in each condition).......................................................................................................................... 112

Table 8.5. Frequencies of the categories of critique discourse found in each form of feedback.114

Table 8.6. A standard topic modeling method LDA was applied to detect the latent topics in the feedback. There were 15 topics in the feedback and the top 5 topic words were listed for each topic....................................................................................................................... 117

Table 8.7. The average and standard deviation of proportions of words in LIWC categories along with p-values associated the corresponding t-test (n = 1000)........................................................................... 122
List of Figures

Figure 1.1. JCPenney’s “Hitler” teakettle........................................................................................................... 2

Figure 2.1. Two alternative designs from a student team were submitted to on Amazon’s Mechanical Turk for A-B testing [Dow, Gerber and Wong 2013].................................................................................. 8

Figure 2.2. The Crowd vs. Crowd design method [Park, Son, Lee and Bae 2013]. Every team competes with other teams for the reward and each design team consists of designers and crowd. Designers first propose design ideas. Crowd who are interested in a designer’s idea can join the designer as a team member. Each team selects and submits one design. The winning team shares the reward........................................................................................................................................... 8

Figure 2.3. (a) When using Fiveseconds.com, a user uploads a picture of her design and asks crowd free-form questions such as “what do you like most about the design?” (b) The user receives free-form responses collected from the crowd. (c) A word cloud generated by the collected responses................................................................................................................................................................. 9

Figure 2.4. Screenshot of IdeaExpander [Wang, Cosley and Fussell 2010]. The system monitors the content of the group conversation and presents pictures to the group.................................................. 11

Figure 2.5. A critique session is conducted in an industrial design studio. A student is presenting her design ideas, and the instructor and other students are offering feedback on her ideas. ......... 12

Figure 2.6. Screenshot of HYDRA-KITCHEN [Fischer et al. 1993]. The "Current Specification" view lists currently selected answers. The "Catalog" view shows previous designs. The "Current Construction" view shows the current design. The "Messages" view automatically provides critic notification messages related to the current design. The critic messages are ranked by the relevance of the critic........................................................................................................................................................................... 13

Figure 2.7. Screenshot of graphical annotation tools provided by sense.us [Heer et al. 2007]. The system enables users to attach both graphical and text annotations................................................................. 15

Figure 2.8. Both word cloud and tag cloud are used to visualize Barack Obama’s speech at the Democratic Convention in 2008. 70% of participants in the study reported that word cloud were more effective than tag cloud [Viégas et al. 2009]. ........................................................................................................... 15

Figure 2.9. A multiple coordinated views presentation of diverse information relating to legal cases [North and Shneiderman 2000]. The system consist of the following views: query results shown in the view in textual form and the view in a graphical visualization, and the selected query results shown in overview and detail views................................................................................................................. 17
Figure 3.1. In (a) a photograph has received two critiques from the community. For each critique, the interface shows the user who wrote it, his reputation score (Callout 2), his rating of the photo (Callout 1), and the rating this critique has received from the community so far (Callout 3). After reading the critique, a user can rate it (Callout 4). If a user selects “write a critique” (Callout 5), the system responds with (b). In (b), the user rates the photo and writes the critique. Once submitted, it would appear with the other critiques of that photograph in (a) .... 21

Figure 3.2. Distribution of percent change from a user’s first 10% of submissions (avg.) to their most recent 10% (avg.) ... 26

Figure 3.3. Distribution of photos receiving different numbers of critiques. 29

Figure 3.4. Distribution of the response time of the critiques. 30

Figure 3.5. Distribution of the total ratings of critiques. 33

Figure 3.6. Percent of users who only write critiques, only receive critiques, or both, as a function of the number of items (critiques and photos) contributed. 34

Figure 3.7. (a) User A critiques user B’s photo, B then critiques A’s photo; (b) User A critiques any photo in the community and someone else in the community critiques A’s photo. .... 35

Figure 3.8. Comparison of rating and response time between reciprocal and non-reciprocal critiques. The data are skewed and therefore analyzed using the Mann-Whitney U test ($N_{\text{recip.}}=1000$, $N_{\text{non-recip.}}=1000$). (a) The median rating of reciprocal and non-reciprocal critiques is 3 and 2, respectively ($p<.001$); (b) the median response time for reciprocal and non-reciprocal critiques is 2.4h and 0.8h, respectively ($p<.001$). 37

Figure 3.9. A visualization of the aggregate pattern of reciprocal critiques in the community. About 30,000 users were sorted in descending order of their reputation scores and divided into 100 groups. Each node represents a group of 300 users. The larger size of a node represents users with higher reputation scores. Each edge indicates that more than ten reciprocal critiques were made in total between any two users in the groups. A majority of reciprocal critiques was made between users with similar reputation scores. Also, users with higher reputation were more active in making reciprocal critiques. 39

Figure 3.10. The logo of the CHI 2011 conference. I submitted this logo to an active online critique community1 and received two critiques. One was “I like the layout and color”, and the other was “the tagline in the logo is difficult to read”. These critiques only have a few sentences, lack in-depth analysis of the design, and offer little insight for consideration. In contrast, I receive a more comprehensive critique on this logo via the implemented approach. (see Testing My Approach for details). 43
Figure 5.1. The interface for configuring the feedback generation process. From this interface, Alice uploads her poster image, configures the desired audience (M/F, USA, 18-35) and selects which feedback is desired. For example, after selecting Goals, Alice enters her communicative goals for the poster. .................................................. 59

Figure 5.2. The user interface offers coordinated views for each type of feedback generated on Alice’s poster. In the Elements feedback, she reviews the list of elements seen by the crowd. .............................................................. 61

Figure 5.3. First Notice feedback. The user views which elements and corresponding regions of the design are first noticed and creates a filter to determine which elements relate to the center markers. ........................................................................................................ 61

Figure 5.4. Impressions feedback. A word cloud represents the impressions and their frequency that the crowd had upon first viewing the design. ................................................................. 63

Figure 5.5. Goals feedback. A set of interactive charts show the distributions of the crowd ratings for each communicative goal. .................................................................................................. 63

Figure 5.6. Guidelines feedback. A similar set of charts shows the distributions of the ratings for each design guideline. The user selects the non-positive ratings to inspect the corresponding explanations and visual markers. .................................................................................. 65

Figure 5.7. Guidelines filtered. The user further explores the feedback by selecting a subset of the visual markers (bottom center), which filters the ratings shown in the chart and the explanations. ........................................................................................................ 65

Figure 6.1. The workflows and related task screens for generating Elements feedback in Voyant. (a) the worker enters at least two elements seen or perceived in the design and then (b) sizes and locates a rectangle to locate each element. .................................................................................. 69

Figure 6.2. The task screen for generating First Notice feedback in Voyant. A worker selects which element is first noticed in the design......................................................................................... 70

Figure 6.3. The workflows and related task screens for generating Impressions feedback in Voyant. (a) a worker enters the first word(s) that come to mind after viewing the design. For each impression entered, (b) the worker locates the most related area of the design and enters rationale.................................................................................................................. 71

Figure 6.4. The task screen for generating Impressions feedback in Voyant. Additional workers vote on the collected impressions and enter rationale ................................................................. 72

Figure 6.5. For Guidelines feedback, a worker learns about the guideline Proximity. .................. 72

Figure 6.6. Related task screens for generating Guidelines feedback. (a) After a worker learn a design guideline (see Figure 6-5), he rates the design on the guideline, and (b) locates the area of
the design that best supports the rating and enters rationale. The workflow for Goals is the same as that for Guidelines. .................................................................73

Figure 7.1. System infrastructure overview. Voyant provides users on-demand access to design feedback from groups of non-experts. The user formulates her feedback request on a design through a web-based interface. The system instantiates the workflow to generate the configured feedback by recruiting workers to perform the set of Describe and Interpret micro-tasks. The system aggregates the task outcomes and allows users to explore the crowd based feedback through visual analytics. .................................................................77

Figure 7.2. One of the initial prototypes of Voyant created after initial discussion with my advisor and teammates.................................................................80

Figure 7.3. Entity relationship diagram is used to visualize the relations among the tables in the database.................................................................82

Figure 7.4. A tree view for folders in Voyant.................................................................83

Figure 7.5. A series of commits were made to Voyant. These commits including editing the feedback request page, displaying feedback generation speed, and fixing bugs with IE. ............84

Figure 8.1. The designs collected for the evaluation included Web designs, a magazine cover, and an illustration.................................................................88

Figure 8.2. The Impressions feedback generated by Voyant on a user’s poster. The user intended it to be perceived as Shakespeare, but was surprised to learn of an unintended interpretation (see “dog” in word cloud). The interface is showing a word cloud representing the impressions formed when the crowd first viewed the design. Selecting an impression (in orange) shows the corresponding explanations and visual markers associating specific regions of the design........90

Figure 8.3. Part of the feedback (Goals) generated by Voyant for a Web design collected from PL20. The participant explored the non-positive ratings and learned that the icons were not being perceived the way he intended and needed refinement. .................................................................91

Figure 8.4. One design goal of the 3D rendering is render teapot in Chinese style. The participant learned that this goal was not well supported. After exploring the visual markers and explanations, she decided to add tea cups beside the teapot and add a dragon pattern to the teapot to better communicate Chinese style. .................................................................91

Figure 8.5. The crowd has opposing views on how well the design meets the guideline. This is useful because the user can see how people interpret the same thing differently, and the user can possibly find a compromise between them.................................................................93

Figure 8.6. The filed study consisted of three phases.................................................................97
Figure 8.7. The designs created during the study. (a) Initial Designs: preliminary designs completed by participants in the first week. (b) Revised Designs: after receiving crowd feedback, participants spent one week revising their initial designs. Each column represents the initial / revised pair of designs for a participant. ................................................................. 98

Figure 8.8. Distribution of ratings of the depth of changes made on the Initial Designs. .............. 98

Figure 8.9. In (a), the first noticed element by the crowd is “building”. In the revision (b), the first noticed element is the event. The feedback helped the participant [PF59] understand how to utilize an image to attract people’s attention more effectively. .................................................................. 101

Figure 8.10. The Impression feedback on the initial and revised designs for the participant [PF54]. (a) The impression word “hard to read” on the initial design prompted the participant to change the image treatment. The revision (b) was no longer perceived as “hard to read” by the crowd. ......................................................................................................................... 103

Figure 8.11. In (a), the design goal “visually appealing” is not well supported. After the participant [PF56] changed the layout and typeface, the average rating on the design goal in (b) was improved from 0.05 (σ = 1.5) to 1.5 (σ = 1.2). The range of the rating scale shown is -3 (low) to 3 (high). ........................................................................................................................................................................ 105

Figure 8.12. The Guidelines feedback on the participant’s designs [PF58]. Based on the feedback, the participant changed the layout and color in (a). The average rating on the guideline alignment in (b) was improved from -1.0 (σ = 2.3) to 1.3 (σ = 1.4). ................................................................................................................................. 107

Figure 8.13. Overall quality ratings from non-experts and experts on the Initial and Revised Designs. ............................................................................................................................................. 108

Figure 8.14. Distribution of the rating changes on Guidelines. ....................................................... 110

Figure 8.15. Distribution of the rating changes on Goals. .............................................................. 110

Figure 8.16. A heatmap visualizes the similarity values between any two topics. The lighter color means the similarity value is higher. The topics in Voyant feedback are more diverse than the topics in free-form feedback. .................................................................................................................................. 119

Figure 9.1. Voyant generated the Impressions feedback on the same design at different times. Frequent impressions such as “colorful” and “juice” in the word clouds remain the same, but less frequent impressions are different in (a) and (b). ......................................................................................................................... 126
Chapter 1
Introduction

Design is a dialectic between goals and possibilities [Stefik et al. 1987]. It pervades everyday activities and generates a wide spectrum of creative, social, and economic values for individuals and society. However, design can be incredibly hard [Edelson 2002; Schön 1992]. It is known to be a highly creative activity that demands problem-solving skills and the ability to reflect upon one's work. The design space is also typically ill-structured, requiring tremendous effort to generate and explore design alternatives. Moreover, an audience's perception of a design can be frequently unpredictable and contextual. Even expert designers with many years of professional experience may have difficulty anticipating the perceptions of an audience. Such challenges are almost inevitable in the design process, but can be addressed through feedback.

Feedback is an indispensable part of the creative process across design domains. Feedback enables people to see their work through the eyes of others and thereby increases understanding of the work and guides further development [Hundhausen et al. 2012]. For example, receiving feedback reveals gaps between what the user intends and what others interpret in a design [Elkins 2012]. Feedback also helps remove creative blocks [Nijstad and Stroebe 2006], inform design choices [Dow et al. 2011], and gain insight into the design [Tohidi et al. 2006].

1.1 An Example of Why Audience Feedback Is Needed

Gaps between intentions and interpretations are not uncommon in design and, if not addressed, can have adverse consequences. In May 2013, JCPenney released a teakettle (see Figure 1-1) designed by a well-known designer Michael Graves, who has created more than 2,000 highly valued designs for major companies. The intent was to create the perception of a
snowman. However, many customers perceived that it resembles Adolf Hitler. As a result, JCPenney made a public apology and discontinued selling the teakettle in stores.

Knowing where the gaps exist is therefore critical for helping users iterate toward solutions that better connect with the audience [Dow et al. 2009]. This dissertation presents a new platform that allows designers to access feedback from a non-expert crowd in order to evaluate how the audience perceives different aspects of the design.

1.2 Existing Options for Receiving Feedback

Feedback is a limited resource. A user currently has several options for receiving feedback on designs. One option is organizing a critique [Elkins 2012], where the comprehensive feedback fosters insight into the design problem, solution approach, and strategies utilized. Though organizing a critique is warranted at checkpoints (e.g. as scheduled by clients or course instructors), it is burdensome for day-to-day decision making. Also, users without a background in design may not have the peer network or desire to organize a formal, face-to-face critique.

A second option is to informally ask peers for feedback, which incurs minimal overhead and can yield quick results. However, the number of feedback requests is limited and the feedback can be biased by friendship or competition [Tohidi, Buxton, Baecker and Sellen 2006]. The
user may also feel uncomfortable showing her design to others due to the fear of criticism [Diehl and Stroebe 1987] or may be concerned about interrupting their work [Fogarty et al. 2005]. A third option is to submit a design to an online community (e.g. Core77) where diverse and rapid feedback is possible. Unfortunately, studies report that little feedback is generated and the quality does not typically exceed judgments such as “I like it” [Willett et al. 2012; Xu and Bailey 2012].

An emerging paradigm for receiving feedback on designs is harnessing the power of crowds. For example, researchers have enabled users to pose design alternatives to a crowd to collect preferences [Dow et al. 2013]. The advantage is that crowd workers offer a scalable, on-demand, and diverse workforce but the disadvantage is that the workers do not typically have design expertise or knowledge of the project. To advance this paradigm, it is critical to explore what feedback on a design would be desirable from a crowd of non-experts, how that feedback could be constructed through a system, and how that feedback can affect users and their design outcomes.

1.3 A Crowd-based System for Generating Feedback

My dissertation presents Voyant, a novel system engaging an online crowd as a simulated audience to collect, aggregate, and present their interpretations of a design to the user. The key innovation is that the system generates structured feedback on visual designs from non-experts. Specifically, Voyant generates five types of feedback: a list of elements seen in the design, the order in which the elements are noticed, the impressions formed when the design is first viewed, and the interpretation of the design relative to guidelines in the domain and the user’s stated communicative goals.

For each feedback type, Voyant provides coordinated views that consist of a visual overview of the crowd’s perceptions and visual markers overlaid on the design. This interactive model enables analysis of the association between the perceptions of a design and the visual elements within it. To use Voyant, the user imports a design image and configures the crowd demographics to specify the target audience. Once generated, the feedback can be utilized to help iterate toward solutions that better connect with the intended audience.
Compared to other options for receiving feedback, Voyant does not require scheduling or use of social capital. It reduces the fear of criticism since the user and crowd workers are unlikely to be able to identify each other [Paulus and Yang 2000]. It eliminates concern about interrupting other’s work by employing paid workers and produces feedback beyond simple judgments by structuring crowd workflows.

1.4 Scope

My dissertation targets visual design because of the growing needs of visual communication skills and the wide availability of graphic design software. The system proposed in this work is therefore most applicable to visual design. The form of visual design can be posters, logos, illustrations, or Web pages created for courses, research, business or personal interests. However, the concept of collecting feedback from a non-expert crowd is applicable to other design domains. For example, in building architecture, architects could receive impressions of the form of a building or interpretation of its functional goals.

The focus of this dissertation is primarily on design feedback, which is only one part of the creative process. The dissertation does not address issues related to other creative activities such as prototyping ideas, managing ideas, or commercializing ideas.

The dissertation focuses on how to utilize a non-expert crowd to generate feedback. The proposed system recruits workers from an online labor market (e.g. Amazon Mechanical Turk), but other crowds such as the employees of a large organization, members of an online community, or students in a large design course could also be used.

The potential audience of the system includes novice and expert designers as well as non-designers who are performing visual design work for personal interests. In other words, any person creating visual designs can use the proposed system to assess how non-experts perceive the designs.

1.5 Contributions

My dissertation work advances knowledge of how to crowdsourcedesign feedback. Specifically, my dissertation makes four central contributions:
Empirical results showing the feasibility of receiving feedback from online crowds. First, I reported user experiences of receiving feedback from a large online design community. The results reveal that little feedback is generated. The feedback quality does not typically exceed judgments such as “I like it” or “I don’t like it”, which offers little benefit to the user. Second, I explored the feasibility of improving online feedback by combining a theoretical framework of design critique with crowdsourcing techniques. The results showed a promising future for crowdsourcing feedback and inspired this dissertation. These details are described in Chapter 3.

A taxonomy of the types of feedback desired from non-experts. I conducted a qualitative study to identify the types of feedback on a design that are desired from non-experts. Semi-structured interviews were conducted to gain insight about users’ current experience of receiving design feedback, the types of feedback desired in the design process, and the desirability of receiving feedback from non-experts. I interviewed both experienced designers and hobbyists who occasionally create visual designs (e.g. data graphics, posters for conferences, or personal Web sites) as part of their work or interests but had no formal training in visual design. The results from the formative study provided a taxonomy that can guide crowd-based systems to generate non-expert feedback on visual designs. These results are described in Chapter 4.

An implementation of a crowd-based system that generates desired design feedback. I designed and implemented a new system to generate structured feedback from a crowd of non-experts. The system decomposes the feedback generation process into sub-tasks suitable for non-experts, aggregates the results, and presents them to the user. An innovation of Voyant is that any user performing visual design work can use it to assess how non-experts perceive her design. Compared to other options for receiving feedback, Voyant does not require scheduling or use of social capital. It also removes evaluation apprehension since the user and crowd are anonymous. It eliminates concern about interrupting other’s work by employing workers, and produces feedback beyond simple judgments by structuring crowd workflows. These details are described in Chapter 5, 6 and 7.
Empirical results of the utility of crowd feedback in the design process. First, I investigated the utility of the feedback generated by a crowd-based system with users and their designs in a lab study. Users reported the feedback helped them identify design problems, gain insight into their designs, and consider their designs from diverse perspectives. The monetary cost of generating the feedback was considered a fair tradeoff for not having to organize critiques or interrupt peers. Second, I examined how the system is utilized throughout design iteration in practice. Most participants reported that Voyant feedback helped them make substantial thematic changes. Third, I examined the content difference between Voyant feedback and free-form feedback, and found that Voyant feedback was more diverse and detailed than free-form feedback. These details are provided in Chapter 8.

My dissertation aims to give users access to a non-expert crowd to receive helpful, affordable, and timely feedback for their visual design. This dissertation could have a large impact on design practice. Crowd feedback integrated into the creative process could allow more design decisions to be based upon data rather than solely on intuition. As a result, crowd feedback systems have the potential to tighten the design cycle and help users iterate toward more effective solutions.
Chapter 2

Literature Review

My work focuses on developing crowd-based systems for generating design feedback. In this chapter I discuss existing work in six key areas that are most relevant to my work and situate the work within the existing body of literature. First, I provide an overview of the types of crowdsourcing techniques proposed for design work. Second, I discuss needs and effects of feedback in the creative process. Third, I consider the theory of feedback in the literature of design education. Fourth, I situate my work in context of other existing approaches for receiving feedback. Fifth, I touch upon how feedback is represented. Finally, I discuss existing empirical studies of online creative communities and the role of feedback in these communities.

2.1 Crowdsourcing in Design

Design is a complex task and researchers have investigated many directions for how a non-expert crowd can aid design. For example, Dow et al. studied how crowd technologies can aid each phase of the design process such as collecting preferences on variations of a design (AB testing) [Dow, Gerber and Wong 2013] (see Figure 2-1). Results showed that students enjoyed receiving feedback from AB tests conducted on crowdsourcing platforms. In the CvC design method, the crowd works with the designer to form a team as part of an open design competition (see Figure 2-2) [Park et al. 2013]. The designer can leverage these team members to collect their preferences on proposed design solutions. Yu and Nickerson showed how crowds and genetic algorithms can be combined to generate concept sketches [Yu and Nickerson 2011]. Others have utilized crowds for rating ideas in innovation competitions [Xu and Bailey 2012], testing interfaces based on task performance [Komarov et al. 2013], and interpreting prior ideas in unanticipated contexts [Faste et al. 2013].
Figure 2.1. Two alternative designs from a student team were submitted to on Amazon’s Mechanical Turk for A-B testing [Dow, Gerber and Wong 2013].

Figure 2.2. The Crowd vs. Crowd design method [Park, Son, Lee and Bae 2013]. Every team competes with other teams for the reward and each design team consists of designers and crowd. Designers first propose design ideas. Crowd who are interested in a designer’s idea can join the designer as a team member. Each team selects and submits one design. The winning team shares the reward.
When using Fiveseconds.com, a user uploads a picture of her design and asks crowd free-form questions such as “what do you like most about the design?” (b) The user receives free-form responses collected from the crowd. (c) A word cloud generated by the collected responses.
My thesis work is original relative to this prior work because it generates *structured feedback* on designs. From user perspective, various types of feedback provided by Voyant is important since prior work solicited only individual preferences which are regarded as the least helpful aspect of design feedback [Feldman 1973]. For example, my approach generates the crowd’s impressions when viewing a design and interpretation of the design according to the user’s goals. For crowd perspective, Voyant structures the feedback generation process into various sub-tasks that can be performed by non-experts. Each sub-task focuses worker attention on specific aspects of a design rather than soliciting holistic evaluations to improve outcomes.

Crowd-based usability Web sites can also be used to collect feedback on designs [Feedbackarmy; Usabilla]. For example, when using Fivesecondtest [Fivesecondtest], a user can pose free-form questions to the crowd about a design and the site returns a word cloud from the responses (see Figure 2-3). In comparison, my system is original because it provides coordinated views for analyzing the association between the crowd’s perception of a design and the visual elements within it. My system also generates specific feedback (e.g. on the user’s goals) not directly provided by these sites. Though the question format of the sites could be used to approximate the questions targeted by my system, the user would need to conceptualize and phrase the questions in a way that yields desired and consistent responses. Finally, this dissertation presents the results of the first study testing the utility of a crowd-based feedback system.

### 2.2 The Creative Need for Design Feedback

From a creative cognition perspective, feedback can foster insight or unblock creativity because the feedback serves as retrieval cues that activate new memory items and thought production [Nijstad and Stroebe 2006; Wang et al. 2011]. This, in turn, can spark new understanding of a design, new solution approaches, or new perspectives on the design problem. For instance, Wang *et al.* proposed IdeaExpander, a Wizard of Oz prototype, to generate visual stimuli based on the conversational content of group brainstorming (see Figure 2-4) and the results showed that exposure to external visual stimuli aids brainstorming groups in producing more diverse ideas [Wang et al. 2010]. Also, Dow *et al.* showed that designers who receive feedback during iterative design produce higher quality outcomes than those who
do not [Dow, Heddleston and Klemmer 2009] and sharing multiple designs with others, due in part to increased communication, improves design exploration and outcomes [Dow, Fortuna, Schwartz, Altringer and Klemmer 2011]. The system proposed in my thesis work strives to allow these benefits to be realized by enabling users to receive feedback as often as desired during the design process.

2.3 Theory of Feedback in Design Education

Feedback in art and design, also referred as critique, provides insights into the artifact and illuminates the technical, cultural and social values reflected in it. In aesthetic and literary theory, critique is the art of discerning the quality of the work, typically through reference to various conventions and mechanisms. First generation of professional art critics included Jonathan Richardson [Richardson 1719], who developed a practical rating system for appraising painting and writing, and Denis Diderot [Goodman 1995], who engaged in the work of judging contemporary visual arts. Published professional criticism is one of the spurs to artistic progress [Barrett 1989]. It can be conceived either as a measured judgment governed by explicit principles, or a personal response without an explicit purpose or thesis [Venturi and Marriott 1964]. On the one hand, according to predetermined principles, art criticism aims to present a universal and reliable assessment of the value of an artwork. On the other hand, art criticism could be the reflection of personal sensibility or taste.
In art and design education, the purpose of critique (e.g. studio and classroom critiques) is to improve students’ abilities to create better artifacts and understand how to better assess creative work (see Figure 2-5). A major finding of previous studies on critique in art and design education is that the quality of critique over quantity is desirable. For example, Sagun and Demirkan have studied how critiques affect students’ design performance in a computer-mediated environment [Adams 2007; Sagun and Demirkan 2009]. Results indicate that quality rather than quantity of critiques determines the success of proposed design solutions. Also, prior research revealed that reciprocity between designers fosters effective communication and collaboration [Dave and Danahy 2000; Dutton 1987]. Critique affords professional and creativity development by providing opportunity for meaningful learning experiences; thereby improving the teachability of the field [Risatti 1987].

In the CSCW community, Stevens et al. have demonstrated the benefits of end user critique for improving information technology artifacts [Stevens and Wiedenhöfer 2006; Wulf et al. 2011]. The authors created interfaces for end users to capture application usage situations and
annotate the specific problem or need. The critiques are integrated and made available to the developers for review.

The quality of critique varies considerably according to where, how and by whom it is given. To improve critique quality, various theoretical frameworks have been proposed to structure the critique process [Feldman 1971]. Students are trained to follow sequential procedures to receive and give criticism. For example, Feldman proposed a well-known four-step process for giving visual art critique: description, analysis, interpretation, and evaluation. 1) Describing visual elements in a design. 2) Analyzing relationships among visual elements in the design. 3) Interpreting the content or meaning derived from the design. 4) Evaluating the design based on the description, analysis, and interpretation. The main idea of this framework is to delay judgment and convert the critique process into sub-tasks. The sequential

![Image of HYDRA-KITCHEN](HYDRA-KITCHEN.png)

Figure 2.6. Screenshot of HYDRA-KITCHEN [Fischer et al. 1993]. The "Current Specification" view lists currently selected answers. The "Catalog" view shows previous designs. The "Current Construction" view shows the current design. The "Messages" view automatically provides critic notification messages related to the current design. The critic messages are ranked by the relevance of the critic.
framework not only fosters high-quality critique but also divides the critique process into sub-tasks, shedding light on mechanisms of aggregating individual efforts to provide comprehensive feedback.

Researchers have developed a typology of critique discourse generated by novice and expert designers [Dannels and Martin 2008]. A wide range of feedback categories in face-to-face critiquing sessions reflects the social and educational context of design. A non-expert crowd may not be able to provide feedback categories such as discussing the design process, precedents and contemporary trends, but it can help users assess how their designs are perceived from an audience perspective, which is an important aspect of design. Crowd feedback systems also enable users to reach a large and diverse audience that would be difficult to obtain in face-to-face settings. The crowd feedback system developed in this thesis enables designers to receive audience feedback on visual designs. Enabling other dimensions of feedback (e.g. process, precedents, and trends) is left for future work. However, in the evaluation, I did examine how crowd feedback compares to expert evaluation and typologies of critique discourse.

2.4 Methods for Receiving Design Feedback

There are at least two approaches for computationally generating feedback on designs. One approach, software critics, is to encode and apply domain-specific knowledge in the form of rules to generate feedback on designs [Fischer et al. 1991; Fischer, Nakakoji, Ostwald, Stahl and Sumner 1993]. Fisher et al. pioneered the development of an embedded critiquing system named HYDRA-kitchen, which is a residential kitchen design environment [Fischer, Nakakoji, Ostwald, Stahl and Sumner 1993] (see Figure 2-6). In HYDRA-kitchen, when a designer changes her design, the system offers the corresponding sets of critics based on the changes. The limitation of this approach is that it cannot consider the unique goals of the user or nuances of the design problem [Schôn 1992]. In contrast, my approach overcomes this limitation because it is based on the coordination of human input. A second approach is to build computational models to predict aesthetics and affect [Zheng et al. 2009], complexity and colorfulness [Reinecke et al. 2013], or perceptual groupings [Rosenholtz et al. 2009] using the visual features of a design. My thesis work generates a broader range of feedback
such as feedback on the user’s communicative goals and established guidelines in the visual domain. Also, because the feedback is synthesized from human input, it can include explanations and inspiration for improvement.

Besides software tools, a user can use social approaches to receive feedback on designs such as organizing critiques, informally asking peers, and participating in online design communities. Critique is the traditional venue for receiving feedback in the creative design domains [Elkins 2012; Feldman 1971]. However, the critique imposes the burden of
organizing discussions, which are typically conducted only at scheduled checkpoints. Users without a background in design may also not have the peer network or desire to organize a critique. A user can ask peers for feedback but may be uncomfortable due to the fear of criticism [Diehl and Stroebe 1987] or concerned about interrupting their work [Fogarty, Hudson, Atkeson, Avrahami, Forlizzi, Kiesler, Lee and Yang 2005]. A design can also be posted to an online community (e.g. Core77 [Core77] or Dribbble [Dribbble]) to receive feedback. Studies of these types of communities, however, report that little feedback is generated and the quality does not typically exceed statements such as “I like it” [Willett, Heer and Agrawala 2012; Xu and Bailey 2012].

In contrast to these social approaches, the feedback in my system can be generated on-demand by any user performing visual design work. It reduces the fear of criticism since the user and crowd workers are unlikely to identify each other and tracing a design to the user should be hard, especially if anonymized, thereby offering anonymity [Paulus and Yang 2000]. It removes concern about interrupting other’s work because it uses paid workers and it uses structured workflows to generate feedback that surpasses simplistic statements.

2.5 Representation of Design Feedback

The HCI and Visualization communities have a long history of developing techniques and systems for efficient representations of various forms of feedback and interaction. The research of graphic annotations and visualizing free-form text seem is most relevant and informative to my work.

Annotation historically refers to the comments written in the margin of paper documents. One of the first scalable annotation support systems was NCSA Mosaic [Andreessen 1993] which allowed users to provide personal annotations of visited web sites. Others could access and build upon the annotations to build collaborative knowledge. A visual annotation system, such as the regional annotations in Wikimapia.org, the anchored conversations of [3], and graphic annotations in sense.us [Heer, Viégas and Wattenberg 2007] (see Figure 2-7) often enable placing visual markers directly within an artifact to support efficient asynchronous discussion. The overlay annotations are often linked to the discussion itself. In general, annotation of work artifacts facilitates organizing information, keeping context, and enhancing memory
In my proposed system, the interface that allows crowds to place visual markers on a design is partly inspired from previous visual annotation systems.

In text visualization, one popular way of representing free-form text (e.g., a document) is as a word cloud [Chuang et al. 2012], which renders a list of descriptive key words or phrases typically sized by raw term frequency. For example, Viégas et al developed a well-known web-based system named “Wordle” that provides an overview of given free-form text via a word cloud with careful design in typography, color, and composition [Viégas, Wattenberg and Feinberg 2009] (see Figure 2-8). Also, in order help users interpret online product reviews, Review Spotlight adopts a word cloud to show the frequency of noun-adjective pairs...
for aggregating reviews [Yatani et al. 2011]. Word cloud will be applied to my proposed system to help users gain an overview of crowd’s perceptions of a design.

System designers often coordinate multiple visualizations, taking advantage of the strengths of each, to create powerful information exploration environments [Baldonado et al. 2000]. Various types of coordination between views have been proposed: brushing and linking, overview and detail view, drill down, and synchronized scrolling. For example, North and Shneiderman developed a system that uses multiple coordinated views to help users explore a data set at multiple levels of abstraction, and from different points of view [North and Shneiderman 2000] (see Figure 2-9). Followed by past work, I propose a comprehensive design for understanding and interacting with feedback on visual designs. Specifically, the system will provide coordinated views that consist of a visual overview of the crowd’s perceptions (e.g. in the form of a word cloud or bar chart) and visual markers overlaid on the design (see Figure 14). The coordinated views enable analysis of the association between the perceptions of a design and the visual elements within it.

2.6 Studies of Q&A and Online Creative Communities

Feedback can be loosely framed as Q&A (what do you think of this design?) and it shares similarities with reviews in product review sites. In [Adamic et al. 2008], the authors studied many aspects of knowledge sharing behavior within Yahoo Q&A. Similarly, Otterbacher showed how review submission time, reviewer reputation, product characteristics, and review content can predict community ratings of ‘helpfulness’ in Amazon.com product reviews [Otterbacher 2009]. Social network metrics have also been leveraged to help predict question quality [Yla R. Tausczik 2011] and question type [Harper et al. 2009] in Q&A. The analysis reported in my thesis borrows from the analyses reported in these prior studies, for example factors such as reputation scores, content attributes, and social metrics are considered in my analysis of an online feedback community.

Researchers have studied other elements of online creative communities in which users collaborate on or share creative artifacts. For example, Luther et al. examined factors that contribute to successful collaboration in a large community of animators [Luther et al. 2010] while Sylvan uncovered predictors of project and social influence in a community formed
around a visual programming environment [Sylvan 2010]. In [Aragon and Williams 2011], the authors revealed the importance of communication in online creative collaborations. My thesis contributes to this corpus of prior work by studying a creative community dedicated to feedback. Feedback can be framed as a form of communication with specific goals and learning how to give and accept constructive comments on creative artifacts.

I explore a type of online feedback community that is unique relative to product review, Q&A, and other creative communities. For example, in a feedback community, users mainly exchange feedback on each other's individual creative work. In addition, my approach is original in that I conducted interviews first to understand users’ expectations of the community and compared them to actual experiences. Finally, I examine the benefit of participation from the ratings of artifacts over time and examine reciprocity.
Chapter 3

Preliminary Work

The primary goal of this chapter is to understand how feedback is provided in the present state of online communities. My work has been carried out to examine current user experiences of receiving online design feedback and explore the possibility to apply crowdsourcing techniques within online feedback practices.

3.1 Analysis of a Large Online Critique Community

Due to the need for design feedback, many online communities have formed in the past decade with shared interest in critiquing creative work in domains such as photography [PhotoSIG], writing [CritiqueCircle], and music [MacJams]. The emergence of online critique communities represents a new opportunity for soliciting feedback from a diverse audience. The goal of my first preliminary work was to understand the quantity, quality, and timeliness of feedback and to understand how it aligned with user expectations within online communities.

I conducted a case study to explore PhotoSIG, one large online community dedicated to critique in the domain of digital photography (see Figure 11). My study examined the benefit of participation (e.g., has the perception of skills improved), the response dynamics (e.g., the quantity, quality, and speed of critiques) and how they relate to expectations, factors predicting critique ratings, and reciprocal interactions. My method consisted of analyzing a large corpus of interaction data and conducting semi-structured interviews with active members (N=12). The statistical analysis explored the content, ratings, and interactions of about 30K users, 500K photos, and 1M critiques spanning a six year period. The interviews were conducted first to help steer our statistical analysis and to uncover motives for participation and users’ expectations of the critiques within the community.
(a) The interface for viewing a photo, initiating a critique of the photo, and browsing or rating the existing critiques. The layout of the interface was compressed for the purpose of explanation.

(b) The interface for writing a critique and rating the photo.

Figure 3.1. In (a) a photograph has received two critiques from the community. For each critique, the interface shows the user who wrote it, his reputation score (Callout 2), his rating of the photo (Callout 1), and the rating this critique has received from the community so far (Callout 3). After reading the critique, a user can rate it (Callout 4). If a user selects “write a critique” (Callout 5), the system responds with (b). In (b), the user rates the photo and writes the critique. Once submitted, it would appear with the other critiques of that photograph in (a).
The results of the case study revealed salient gaps between users’ expectations and their actual experiences in terms of the quality, quantity and timeliness of feedback [Xu and Bailey 2012]. A majority of online feedback do not go beyond spontaneous judgments such as “I like it” or “I don’t like it”, which offers little learning value to the user. Also, I formulated recommendations for design of systems that support community-based critique of creative work.

3.1.1 Feedback Community Studied

I studied a large online community of users with a shared passion for digital photography; PhotoSIG [PhotoSIG]. In the community, users share photographs, feedback on each other’s work, and try to improve their content and creative skills. The community attracts a diverse audience from amateurs to established professionals. Registration is required for participating and both free and paid memberships are offered. Free membership allows participation in all aspects of the community but the number of photographs is limited to one submission every three days. Paid membership increases this limit to three submissions every three days.

As shown in Figure 3-1, when a photo is shared with the community, any user other than the owner can write a critique and rate the photo, or rate the critiques posted by others. A photo is rated on a 6-point scale ranging from three thumbs down (-3) to three thumbs up (+3) without a neutral (0) rating. The total rating of a photo is the sum of the individual ratings. A critique can be rated as helpful (+1) or unhelpful (-1), with the default being no opinion, and the total rating of a critique is the sum of the individual ratings. As in [Mamykina et al. 2011], user activity is incentivized through a reputation system. A user’s reputation is the sum of the ratings of the photos and critiques she submitted and is shown with her contributions. Extra points are awarded for behaviors deemed especially beneficial; e.g., being among the first three users to critique a new photograph.

I chose to study this community for three reasons. First, promoting critique has always been the mission and central design goal of this community, reflecting the importance of critique in design. Second, the community has a large, active user base, the rating methods are mature, and there is an open trace of interaction that can be analyzed. Finally, because the process of
critique supported in this community is general, our results and implications should be of interest to similar communities in other design domains.

3.1.2 Research Questions and Methodology

My study was designed to understand critique within a large online community and centered on these questions:

- What are users’ motives for participating in the online community and what benefits do they receive? How does participation (e.g. writing and receiving critiques) relate to the ratings of their creative artifacts, if at all?

- What are user expectations for receiving critiques online? For example, how many critiques do users want to receive, how quickly do they want to receive the critiques, and what makes an effective critique?

- What are the patterns of critique interaction? For example, what is the level of reciprocity and how important is reciprocity from the user’s perspective?

My study used a mixed methods approach, consisting of quantitative analysis of interaction data and semi-structured interviews with users. Publicly available data was collected from the community using a custom extractor. I analyzed data from 2004 to 2009, as the community’s rating system has been stable since 2004. My analysis comprised 32,029 users, 526,561 photos, and 1,249,163 critiques. The data set includes ratings of each photo, critique content, ratings of each critique, reputation scores, and activity timestamps. Table 3-1 summarizes the data set. My analysis examined relationships between users’ photo ratings and participation in the community, differences between user expectations and experiences, and patterns of critique interaction.
To understand users’ expectations of the community and help interpret the quantitative results, I conducted semi-structured interviews with twelve active users (two female) recruited via electronic postings in the community. Three were professional photographers while the others were amateurs. Their experience in the community ranged from 1 to 8 years and their reputation scores ranged from 250 to 5,600 points (all scores ranged from -53 to 110,301 points). The type of content contributed by the participants included nature, model, fine art, and architecture photographs.

Twelve prepared questions were asked during an interview, but additional questions were asked to pursue unexpected or interesting points raised by the participants. Table 3-2 shows the themes covered and a sample of the questions asked within each one. Interviews were conducted by phone (10) or instant messenger (2), whichever participants preferred. Interviews lasted between 60 and 90 min and remuneration was a $20 gift card. The phone interviews were recorded and transcribed and combined with the chat logs. Following [Glaser

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>μ</th>
<th>σ</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of photos contributed per user</td>
<td>5</td>
<td>16.5</td>
<td>7.0</td>
<td>[0, 2090]</td>
</tr>
<tr>
<td>Photo ratings</td>
<td>3</td>
<td>4.2</td>
<td>4.8</td>
<td>[-8, 110]</td>
</tr>
<tr>
<td>Number of critiques received per photo</td>
<td>2</td>
<td>2.4</td>
<td>2.2</td>
<td>[0, 903]</td>
</tr>
<tr>
<td>Number of critiques written per user</td>
<td>6</td>
<td>50.5</td>
<td>275.7</td>
<td>[0, 26550]</td>
</tr>
<tr>
<td>Critique ratings</td>
<td>2</td>
<td>2.7</td>
<td>2.7</td>
<td>[-106, 117]</td>
</tr>
<tr>
<td>Critique word length</td>
<td>45</td>
<td>54.0</td>
<td>34.9</td>
<td>[1, 1956]</td>
</tr>
<tr>
<td>Reputation scores</td>
<td>27</td>
<td>174.6</td>
<td>866.2</td>
<td>[-53, 110,301]</td>
</tr>
</tbody>
</table>

Table 3.1. Descriptives of user activity in our data set.
Context and Motivation

Can you briefly describe a recent experience of either receiving or contributing a critique in the community? What was the effect of the critique?

What is your motivation for participating in this community? What is the benefit you hope to receive?

Writing, Receiving, and Reading Critiques

For what reasons do you critique photos posted in the community? How do you choose which photos to critique?

How often do you read the critiques of other users’ photos? How does reading these critiques benefit you as a photographer?

How many critiques do you prefer when you submit a photo and how quickly do you want to receive the critiques?

How do you evaluate the quality of the critiques and how do you perceive the ratings assigned by others?

User Experience

What is your experience with having your photos and critiques rated by the community? How important are these ratings?

What do you believe are the main strengths and weaknesses of participating in the community?

Table 3.2. The main themes and a sample of the questions asked within each theme for the semi-structured interviews.

and Strauss 1967], the data was coded to derive common responses.

3.1.3 Benefits of Online Feedback

Users reported their main motivation for participating in the community is to gain diverse feedback on their creative work and to enhance their creative skills (11 users). This is unique relative to the motivations of fun and ideology that drive contributions for other types of user-
generated content sites (e.g. Wikipedia) [Nov 2007]. For example, one user was an amateur photographer who has actively participated in the community for two years. He reported that the largest benefit of participating is the opportunity to gain feedback:

“Wandering alone with my camera is relaxing and rejuvenating to me. Getting feedback on my work is icing on the cake. I enjoy feedback from other members. If I've looked over something, or if an image could be better, etc. Basically, it helps as I am not always as critical of my own work as I should be. Sometimes, I miss things; people point it out to me. It is a good driving force to keep going.” [P5]

Other reported benefits include building and maintaining relationships with other photographers, establishing reputation in the field, and advertising photographic services (8 users). These benefits reaffirm findings from studies of other online communities [Butler et al. 2001; Cook et al. 2009; Diakopoulos and Naaman 2011; Lampe et al. 2010]. For example, another user was a professional photographer who runs an online photography school. His motives included not only honing his craft but also promoting his business:

Figure 3.2. Distribution of percent change from a user’s first 10% of submissions (avg.) to their most recent 10% (avg.).
“I am running an online photography school called...and I intentionally post some photos and critiques for exposure. If people like my work or my critiques, they would like to join my school.” [P2]

Since the primary motivation for participating in the community is to improve photography skills, I examined the change in ratings of a user’s photos from when they joined to their most recent contributions. I used the average photo rating within a given time window as a proxy for a user’s photography skill. I computed the average rating of the first 10% and the most recent 10% of each user’s photograph submissions. Only those users with a minimum of ten photos were considered (N=10,542) given our percent cutoffs, but analysis with other percent cutoffs yielded similar results. A t-test showed the average ratings of a user’s photos improved from the first ($\mu=3.8$, $\sigma=2.6$) to the most recent submissions ($\mu=18.5$, $\sigma=40.7$; $p<.001$).

Figure 3-2 shows the distribution of the percent change in the average rating from the first set to the most recent set of a user’s photographs. Reflecting the prior result, the graph shows that a majority of users (56%) received the same or higher ratings for their photos over time, indicating that participation was helping to improve perceived skill. The figure also shows that a minority (44%) of users’ ratings decreased. Though the true cause of this effect is unknown, it may be that the skill of some users was not progressing.

To gain further insight, a stepwise regression was applied to assess how different variables predict the change in photo ratings. Table 3-3 summarizes the independent variables considered and the regression results. The results show that receiving more critiques with higher ratings relates to an increase in photo rating, implying that the quantity and quality of critiques received has a positive correlation with the development of users’ skills.
In contrast, the duration of participation did not show a measurable effect on the change in photo rating. And, interestingly, writing critiques of other user’s work did not relate to the change in photo rating. This was somewhat surprising given that evaluating the work of others can build one’s own design knowledge [Barrett 1988; Lundstrom and Baker 2009; Reily et al. 2009]. One explanation is that users are not effectively instructed on how to perform effective critique and therefore may not be gaining the full benefit of performing critical analysis.

### 3.1.4 Users’ Expectations of Online Feedback

I investigated users’ expectations of online feedback, how their expectations align with actual experiences, and what design enhancements might close the gaps where they exist.

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>p</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>User’s initial Photo Rating*</td>
<td>-0.343</td>
<td>0.084</td>
<td>&lt;.001</td>
<td>0.30</td>
</tr>
<tr>
<td>Num. of Critiques Received*</td>
<td>0.170</td>
<td>0.058</td>
<td>&lt;.001</td>
<td>0.45</td>
</tr>
<tr>
<td>Avg. Critique Rating*</td>
<td>0.154</td>
<td>0.042</td>
<td>&lt;.001</td>
<td>0.51</td>
</tr>
<tr>
<td>Num. of Photos Submitted*</td>
<td>0.074</td>
<td>0.055</td>
<td>&lt;.05</td>
<td>0.52</td>
</tr>
<tr>
<td>Num. of Critiques Submitted</td>
<td>0.046</td>
<td>0.045</td>
<td>0.304</td>
<td>0.52</td>
</tr>
<tr>
<td>Participation Time</td>
<td>-0.004</td>
<td>0.009</td>
<td>0.646</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 3.3. Stepwise regression was applied to test how different variables of participation predict the change of a user’s photo rating. Four of the six variables tested were significant. The coefficients are from the final model while the R² values are from each regression step. The model included the initial photo rating which is inversely associated with the change in photo rating. This may reflect a regression-to-the-mean effect.
There is a mismatch between the number of critiques a user wants to receive and the number s/he actually receives. In the interviews, I found that users want to receive at least four or five critiques for a photo (10 users), as they felt this offered the minimum feedback and perspective necessary for understanding how to move the work forward. The distribution of the number of critiques per photo is shown in Figure 3-3. While some photos receive many critiques, the vast majority (over 80%) receive less than four. The median number of critiques per photo is two. However, only a small percent (3.9%) receive no critique, which may be explained partly by the reputation incentive for contributing critiques of photos which have less than three critiques at the time.

To understand how it might be possible to encourage more critiques, I inquired about how users currently select photos to critique. From the interviews, users reported they typically select photos by browsing the area of the main page showing top-rated or recent photos. From this set, they look for photos matching personal interest or those submitted by a user with whom they had prior interactions. One solution for fostering more critiques is therefore to personalize this area of the main page, increasing the concentration of photos that may be of interest to the user. For example, filter or visually emphasize photos submitted by those with whom the user has an interaction history or photos matching an interest profile specified by
the user. Another option is to expand the incentive for submitting the first critiques of a photo (e.g., raise the cutoff from three to six and increase the points awarded to the reputation score).

**Content**

Interviews revealed that users expect to receive constructive critiques which contain specific technical evaluation (e.g. related to composition, color, and depth-of-field) along with corresponding solutions, in order to facilitate the learning process (12 users). For example, a user said:

“Constructive criticism with solution is essentially the best criticism. This is a golden rule. If you want to tell me something wrong about my photo, tell me how I can fix it or how to make it different. Simply saying ‘I like the photo’ or ‘I don’t like the photo’ either way it is worthless ... It didn’t make me a better photographer.” [P4]

More interestingly, several users preferred those critiquing their photos to not only tell them what to do but also “show me how to do it” (3 users). For example, a user reported that one user critiquing his photo would often revise his posted photo offline and send an email to show him the result:
“I even got some friends go into their emails and send me the picture to show me how to do it. I had that experience, which was really nice. This is pretty demanding of anybody’s time. So I don’t beg for it.” [P7]

Though ‘acting’ on a photo was appreciated, the cost of acting on photos offline is high and disconnected from the online discussion. An enhancement would be to allow users to apply common technical operations such as crops and color balancing directly to a photo as part of the critique.

Critiques with positive expression can help users develop a sense of membership and further motivate them to engage in the community [Choi et al. 2010]. But new users may be not accustomed to follow this norm. As one user stated:

“If someone put a three or two thumbs down critique, I would like to read it to see what made this person be so aggressive. Most always they will have a small number (user reputation rating) beside their name. Like mine has the 21,000 (user reputation rating). Implying that they are beginner that also implying that they are newcomer to the community.” [P1]

I therefore examined the relationship between reputation score and the percent of positive and negative emotion words written in a critique. This analysis was performed using Pennebaker’s Linguistic Inquiry and Word Count (LIWC). LIWC counts the number of words in a source text that match the words within different linguistic categories. The positive category includes 406 words such as fantastic, love, and nice while the negative category contains 499 words such as hurt, ugly, and nasty [Pennebaker et al.]. A weak positive correlation ($r = 0.21; p<.001$) was found between reputation score and percent of positive emotion words, while a small negative correlation ($r = -0.10; p<.001$) was found between reputation score and the percent of negative emotion words. Users with higher reputation scores tend to use more positive words than users with lower scores. Similar to [Choi, Alexander, Kraut and Levine 2010], system designers should therefore consider ‘socializing’ novices to help them understand how to write effective critiques consistent with the norms of the community.

**Response Time**
From the interviews, users want to receive most critiques within 24 hours after submitting a photo (9 users). Figure 3-4 shows that a photo accumulates about 60% of all its critiques in the first two hours. Although the majority of users were satisfied with the response time, several users wanted to receive critiques sooner (3 users). These users often posted work-in-progress photos and they hoped to get “instant” critiques to help accelerate the working process:

“I expect to receive critiques in fifteen minutes. But sometimes I have to wait for five or six hours to get critiques. Most time I submit photos in the working progress. Especially the photos I have questions about. I post them to get feedback. Because if there is something fundamentally flaw in the photograph, then I can face that it does not work. Or there are small minor details I may miss. I can go back to correct them.” [P10]

To meet these users’ needs, a design enhancement might allow users to assign the label of “work-in-progress” to their photos to notify others that quick, high-level feedback would be appreciated. Other design alternatives to hasten response time include the use of reputation boosts, community guidelines, and active moderation [Mamykina, Manoim, Mittal, Hripcsak and Hartmann 2011].

Revisiting Feedback

Several users reported that they often access prior critiques in order to engage in reflective learning or to reflect on their own experiences as a photographer (5 users). High-quality critiques may attract repeated access from users.

“I do occasionally revisit critiques. If I felt as if the person (the person’s critiques) had an outstanding amount of knowledge, or something particularly worthwhile for me to concern myself with, I'll go back” [P9]

However, the current interface does not offer mechanisms for organizing critiques based on different attributes (e.g. words of encouragement, technical suggestions, or social connections). Designing and including such interaction could better support users’ reflective learning activities.
Who Makes the Feedback

Users reported they prefer to receive critiques from those with similar or more experience than themselves (7 users):

“The person critiquing, unfortunately, does matter. If they're photographically novice, they'll be less prone to give me real technical advice. They don't have a lot of advice to impart... I treasure advice given to me by people who are better than I am.” [P8]

This contrasts with [Reily, Finnerty and Terveen 2009] which found students in an intro programming course prefer to receive critiques of their code from those with less experience. One explanation may be that the students perceived they would receive a better rating if the critique was written by someone with less experience, whereas our users are mainly interested in improving their creative skill. Users also felt that critiques written by those familiar to them were easier to understand than critiques written by those unfamiliar (2 users):

“Sometimes it is more difficult to evaluate critiques from strangers. You don’t have history and you don’t have the same vocabulary.” [P7]

Predicting Feedback Quality
Users felt most critiques received were helpful to a degree, but only a small fraction offered deep insights into the work (8 users). This is reflected in the distribution of the total ratings of critiques shown in Figure 3-5. For example, the figure shows only a few critiques received very high total ratings while 92.2% of critiques received a rating of at least one. Among these critiques, 97.4% received only helpful ratings, 0.3% only unhelpful, and 2.3% received both.

Users also felt that the score of a critique reflects the degree of community consensus around the statements made in it:

“Someone gives a good critique, but I may not like what they have to say. But if it’s followed up by 10 more people agreeing with the critique, then I may tend to agree that they are correct.” [P2]

To gain insight about what contributes to critique quality, I performed a linear regression of the critique score (total rating) with factors identified as relevant to quality. Our analysis uses critique score as a proxy for quality. From the interviews and prior work, I identified three categories of factors; content, critique context, and social attributes. For content, I counted the
number of terms in the content that matched terms found in a comprehensive photographic dictionary, including crop, composition, lighting, exposure, and depth-of-field [Guy 2008]. I also counted the number of positive and negative emotion words using the LIWC [Pennebaker, Francis and Booth]. Response time, word length, and rating of the associated photo served as the contextual factors. For social attributes, I used the reputation score and social connectedness of the user writing the critique [Sylvan 2010; Yla R. Tausczik 2011]. Social connectedness was measured by the betweenness centrality of the critic’s node in the directed graph where an edge in the graph is defined as a critique given from one user to another.

To increase robustness [Otterbacher 2009], 10K critiques with a total rating of at least ten were randomly selected from all of the critiques for analysis. The factors considered and results of the regression are summarized in Table 4. Similar to [Adamic, Zhang, Bakshy and Ackerman 2008; Otterbacher 2009; Yla R. Tausczik 2011], I found that message length, reputation score, and social connectedness are predictors of critique quality. In addition, I found that critiques posted earlier with more technical and positive emotion words receive higher scores.

Interestingly, critiques posted to a photo soon after it is submitted receive higher ratings than those posted later. One explanation is that, unlike product reviews [Otterbacher 2009], the
community appreciates critiques more for their timeliness than for their content, especially for photos assumed to be work-in-progress. Almost all users (99.6%) rated others’ critiques of their own work and a considerable portion (31.9%) of critique ratings were given by the owners of the corresponding photos. Another possible explanation is that later critiques offer fewer new insights relative to the earlier critiques and therefore receive lower ratings. Because of the small correlation between the critique response time and the number of individual ratings it received ($r = 0.10; p< .001$), it is unlikely that later critiques receiving fewer ratings accounts for the difference in the total critique rating.

3.1.5 Feedback Reciprocity

Designers often rely upon one another for critiques, a form of reciprocity [Alexander 1987; Dutton 1987]. This is enabled in the community studied by allowing users to both write and receive critiques. In fact, a few users reported that writing critiques was an effective way of scaffolding their participation. The distribution of Figure 3-6 shows that users may begin by submitting only photos or critiques but quickly become engaged in both activities. This motivated us to examine both direct and generalized reciprocity in the community.

**Direct Reciprocity**

Direct reciprocity refers to dyadic relations between users making critiques (see Figure 3-7a). From the interviews, most users reported making reciprocal critiques (11 users):

“It is a polite thing to do, if they critique your works it's nice that you do the same for them.” [P11]

Reciprocal critiques can help build relationships. Many users reported that they have the experience of creating connections by critiquing each other’s work (11 users):

“Sometimes I critiqued somebody and they critiqued on mine. There is a part of it becomes an unspoken relationship. You say something to me and I say something back to you. And eventually we get on the track. We understand each other very well. I believe what happened it is the best when you recognize something in the work of somebody else
that resonates with you that form the possibility of communication and mutual learning.”

Users reported that they would pay special attention to the people they know and tend to give them more serious critiques (7 users):

“So many critiques become personal. That is with the people that your response to and for which there are some sort of mutual respect. You tend to spend more time and care in talking about their pictures and in reading or hearing from them about your own.” [P12]

“It depends on who is on critique. If it is a good friend, I would be likely to give him/her an honest opinion. If I see something could be improved and some other techniques might be more successful, that kind of thing. Otherwise, you know, the critiques would be pretty light. For example, ‘I like the photo’, ‘It gets the great color’, and ‘the subject is interesting’ that kind of thing.” [P3]
Surprisingly, however, only 18.1% of all critiques in the community are reciprocal and the degree of direct reciprocity\(^1\) in the community is only 0.14, which is lower than other social networks such as blogs (0.4), and email networks (0.231) [Garlaschelli and Loffredo 2004]. The rating of reciprocal critiques was higher than the rating of non-reciprocal critiques (Figure 3-8); however, the response time of reciprocal critiques (median=2.4 h) is longer than that of non-reciprocal critiques (median=0.8 h). One possible explanation is that the community did not provide an immediate way for users to reciprocate. For example, users are not notified when those who have previously critiqued their work contribute a new photo. If a user wants to reciprocate, she must access her own critiques, identify the authors, and then visit the authors’ photo sets. Additionally, there was no difference detected between reciprocal and non-reciprocal critiques in the photo rating carried by the critique. This indicates that users do not adjust their rating according to whether this is a reciprocal critique.

The relationship between direct reciprocity and reputation score is visualized in Figure 3-9. It shows that similarity in reputation score correlates with making reciprocal critiques. One user aptly explains this observed behavior:

“One can learn from someone closer in experience. When it's a wide gap, the learning curve is different. The expert gets tired of the same old questions and giving the same old answers. The novices or lesser knowledge have to have somewhere to start to ask the question. The closer the experience, the closer the communication.” [P11]

This behavior reflects homophily [McPherson et al. 2001], where increased similarity in users’ experience increases the likelihood of communication and reciprocity.

**Generalized Reciprocity**

Generalized reciprocity (see Figure 3-7b) is an exchange among individuals without the expectation of immediate or specific returns but with the expectation that the benefit will balance out over time [Alexander 1987]. In the interviews, most users believed the community should balance the effort of writing critiques with the benefit of receiving critiques (11 users):

---

\(^1\)The degree of direct reciprocity in the community is calculated by the link reciprocity which considers both the ratio of the number of links pointing in both directions to the total number of links and the network density.
“It's basically the adage of "you get what you give." If you never make critiques, your popularity in the community should likely reflect this. This is how it works for about 99 percent of us, at least.” [P8]

“I feel that on any photography site there should be a good family community where we all work and strive together to help each other learn.” [P1]
However, the correlation coefficient between the number of critiques a user writes and receives was only 0.19 ($p<.001$), where the number of photos contributed was used as the control variable. The weak correlation indicates that users often write more critiques than they receive and that a more balanced allocation would be welcome.

### 3.1.6 Design Implications for Online Feedback Communities

#### Structure Feedback

Many users reported wanting their designs to be critiqued from multiple technical perspectives, such as composition, color, and depth-of-field, but there is no direct way of expressing this preference. An alternative design would be to allow the owner of an artifact to specify which technical perspectives they would like to receive feedback on and this information would be viewable in the interface for writing a critique. For example, the system can provide a comprehensive list of technical perspectives from which the owner can choose. As part of writing a critique, a user would have to select which of those elements are being addressed. This direction requires careful scrutiny because although specific requests for information can elicit greater community responsiveness [Burke et al. 2010], it may also increase the micro-costs of writing critiques and affect participation [Benkler 2006]. This opens an interesting research question about how to steer community responses towards the preferences of the owner of an artifact without compromising participation.

#### Foster Direct and Generalized Reciprocity

The value of both forms of reciprocity was revealed in our interviews but little reciprocity was observed in the community overall. To catalyze direct reciprocity, one solution is to make visible the ongoing reciprocal relationships and current status, e.g., identify who has critiqued your work, number of critiques made, and how many times you have critiqued their work. Systems could be even more socially aware by suggesting photos to be critiqued based on factors such as similarity of expertise between users and their prior interaction history. Expertise could be approximated using techniques found in [Zhang et al. 2007] or, more crudely, using only the reputation score. Suggesting photos to be critiqued is similar to
matching users with tasks that need to be performed [Cosley et al. 2007], but the purpose would be to cultivate interpersonal relationships in addition to generating the critiques. To catalyze generalized reciprocity, one solution is to weigh the number of critiques and ratings of those critiques more favorably in the computation of reputation score. For those who often contribute critiques but receive few in return, another solution would be to give their work more exposure in the public areas of the site, e.g. display them on the main page longer, which may generate more critiques. A third solution would be to make visible the balance of critiques written vs. received for each user. This may cause others to adjust their decisions in regards to whose works they would critique.

**Support ‘Acting’ on the Artifact Being Critiqued**

In a critique, users often recommend applying various technical operators to the artifact, e.g., cropping or color balancing in photography. However, these operators are almost always described in words and it is up to the owner and community members to mentally simulate how these operations affect the artifact. To increase the communication effectiveness of a critique, one solution would be to allow an artifact to be directly edited as part of writing the critique. The edit operations could be captured, represented, and displayed with their written critique. This concept may look similar to the structured comment space created in [Willett et al. 2011], but the visual vocabulary of operators would be specific to photography or other design domain. In addition, when a user responds (replies) to an existing critique, the interface could allow the user to build on or modify the operators applied in the prior critique. This would simulate the face-to-face critique environment in which users can build on each other’s ideas.

**Manage Variations and Revisions**

Users sometimes post variations of the same content in order to receive critiques and guide their design choices. I also found that users would revise or retake photos based on prior critiques and resubmit them. The problem is that they are unable to create associations between variations of the same content or revisions of their work; rather each item is viewed and managed as a separate artifact in the system.
Researchers have created interfaces for managing variations and alternatives (e.g. see [Hailpern et al. 2007]), but these interfaces target single-user applications rather than online communities. For the type of community I studied, variations could be supported by allowing a user to associate multiple artifacts as a ‘variation set’ indicating to the system these should be shown together when viewed and critiqued by others. For revisions, the system could allow a user to associate a new artifact with a prior version. When viewed, the associations could be navigated to see the evolution of the content and corresponding critiques. The benefit would be that a user could manage her design space more effectively and others could better understand the evolution of her work [Tohidi et al. 2006].

**Organize and Revisit Feedback**

Users may receive hundreds of critiques while participating in the community, yet there is no means for organizing the critiques based on the knowledge or insights contained. In order to enable more effective reuse of critiques, systems could allow users to create and apply their own personal or community-derived tags to the content of the critiques, create associations between critiques of different artifacts that make similar points, and aggregate parts of critiques that share the same tags to support review and reflection [Gilbert and Karahalios 2010; Yatani, Novati, Trusty and Truong 2011].

**3.2 Crowdsourcing Design Feedback**

As pointed out in the analysis of feedback community, posting designs to online communities is a convenient way for designers to obtain critiques of their work. However, there is no guarantee of the quantity or quality of critiques received. Also, online communities allow users to comment on designs in an unstructured manner, which is inconsistent with the theory of critique in art and design education [Feldman 1971]. In addition, many users in online communities have insufficient skills or limited time and their individual efforts may not be able to generate a complete critique.

Consequently, online feedback typically does not go beyond simplistic statements of “I like it” or “I don’t like it”, which offers little benefit to the designer (see Figure 3-10). My second preliminary work attempts to improve the quality of online feedback by integrating a
A theoretical framework of design critique with crowdsourcing techniques. The main concept of my preliminary prototype was to convert the critique process into a series of sub-tasks using the sequential critiquing framework and selectively aggregate the results of the individual efforts form comprehensive feedback.

The goal of my second preliminary work is to explore the feasibility of generating quality design feedback via crowdsourcing techniques. A main reason of using crowdsourcing is that iterative tasks divided by the sequential critiquing framework could be efficiently performed by paid crowd workers. In contrast, it is difficult to directly apply a sequential framework in the context of online design communities. Although the framework can convert the critique process into sub-tasks, requesting community members to follow specific instructions strictly and work on the sub-tasks iteratively is not consistent with how the community functions, and thus participation would be affected [Benkler 2006]. I aim to propose a crowdsourcing model to generate design feedback.

### 3.2.1 Sequential Process for Feedback

A major finding of previous studies on critique in the art and design education is that the quality of critique over quantity is desirable [Barrett 1989; Graham 2003]. Feldman [Feldman 1971; Feldman 1994] developed a well-known sequential approach to ensure the quality of critique. It is a four-step process including description, analysis, interpretation, and judgment. The main idea of this framework is to delay judgment and convert the critique process into sub-tasks. Teo and Chai [Teo and Chai 2009] followed the Feldman’s idea and applied the
framework practically in design studios. The simplified framework has four sub-activities: 1) Describing the purpose and audience of the design. 2) Evaluating strengths of the design. 3) Evaluating weaknesses of the design. 4) Summarizing the statements in the preceding steps.

The sequential framework not only foresters quality critique but also divides the critique process into sub-tasks, shedding light on mechanisms of aggregating individual efforts to make a complete critique.

### 3.2.2 Crowdsourcing Model for Online Feedback

My approach attempts to provide a low-cost way of accessing quality critiques quickly. The main idea of my approach is to convert the critique process into a series of sub-tasks according to the sequential critiquing framework and aggregate individual efforts to generate a complete critique via a crowdsourced platform.

It is difficult to directly apply the above framework in the context of online critique communities. Although this framework can convert the critique process into sub-tasks, requesting community members to follow specific instructions strictly and work on the sub-tasks iteratively would dramatically increase the micro-costs of contributing critiques and affect participation [Benkler 2006]. Also, implementing the framework would also require large changes to the underlying system software driving the community.

Fortunately, iterative tasks divided by the framework could be efficiently performed in a for-pay human computation system like Amazon Mechanical Turk (MTurk) [Little et al. 2010; Little et al. 2009]. I apply the above framework to MTurk by converting the critique process into a series of sub-tasks and designing various voting schemes to ensure the responses meet the requirements. The framework is implemented in JavaScript and the API provided by Turkit [Little, Chilton, Goldman and Miller. 2009].

**Description**

This step is used to gather information and help later workers (Turkers) make in-depth evaluation in the next steps. In order to avoid the increased cost of requesting critique, we do not require a designer to elaborate his/her own design intentions. Our approach instructs
workers to accomplish this task, since a designer’s intentions can be described by others [Teo and Chai 2009].

Specifically, workers are required to view the design and answer the following questions: 1) What do you believe are the designer’s intentions? 2) What is the target audience of the design? After the first worker answers these questions, subsequent workers will be instructed to improve the initial description by either revising the writing of the description or adding new ideas to the description. Since workers may have different describing styles and diverse aesthetic experiences, we adopt the voting scheme to prevent the introduction of errors [Little, Chilton, Goldman and Miller. 2009]. Every improvement made by a worker will be voted by at least two other workers. If both of the workers vote for the improvement, it will be kept for the next iteration; if both of them vote against the improvement, it will be rejected; and if the two votes disagree on the improvement, then another vote will be made by another worker to decide whether the improvement will be kept.

**Strength**

In the step, workers are first required to view both the design and its description generated by previous workers. Then, workers are required to point out which element in the design supports the designer’s intention and why this element reinforces the design concept. In other words, workers are instructed to keep a positive attitude, find the strengths of the design, and relate the strengths to the designer’s intent. We create five separate assignments in this step and each assignment requires a worker to provide at least one strength of the design. As a result, the five assignments will collect at least five strengths of the design.

**Weakness**

Similar to the prior step, this step “Weakness” first asks workers to view the design and the corresponding description. Then the workers are instructed to identify elements in the design that do not effectively realize or support the designer’s intent. The workers are required to reveal the weaknesses of the design and offer corresponding suggestions. The steps “weakness” and “strength” can be performed in parallel.
Summarize

The final step consists of two stages: selection and summarization. In the first stage, workers are required to view the design and its corresponding description, and then vote on each strength and weakness to indicate whether they agree with the statement. Multiple independent agreements could suggest that the statement is correct. Our approach aggregates all of the votes and ranks these statements. The top three strengths and weaknesses will be selected for the next stage.

During the second stage, workers are asked to summarize the strengths and the weaknesses into one paragraph respectively. Multiple versions of summaries are generated, and only one summary of the strengths and the weaknesses will be selected based on the subsequent workers’ votes. The final version of the critique consists of three paragraphs: description, strength, and weakness.

3.2.3 Testing My Approach

The goal of my study was to test the feasibility of crowdsourcing design feedback. As the sequential framework was originally developed for critiquing visual designs, we use the logo of the CHI 2011 conference, a representative visual design, to test our approach. The received feedback systematically analyzes the logo design:

**Description.** “The logo has text in dark blue-green, faded blue-green and yellow. The dark color showing "CHI" is the first recognized and that shows the importance of the organization. The faded blue-green "20" of 2011 overlaps the dark blue-green C which changes the CHI into a unique logo to represent just the conference. The "11" of 2011 is smaller, yellow and in a different font compared to the same font being used for the rest of the logo. The entire logo represents integration and differences in color and boldness make the more important information easier to remember - "CHI" and "11". The target audiences are likely to be those who know what CHI is, such as members of CHI or organizations and people that deal with CHI because they work in a similar field. The logo promotes the conference to the target audience in the hopes they will attend.”
**Strength.** “The logo is very popping. The 20HI stood out first, after it engaged me it I saw the CHI and then the 11, the next thing i looked at was the Vancouver, BC. It flows well and is intriguing to the eyes. Another thing that caught my eye is that the 20HI kinda plays a trick with your eyes make me think 2011. I think it looks great the way it is.”

**Weakness.** “While the logo is overall well done, better clarity could be achieved through the following adaptations: First, the light blue is a weak color in contrast to the teal-- changing all of the light blue text to pale orange would be a nice substitute. I would remove the italics on the "11," increase the font size, and change it to a brown or tan. Also, it is difficult to read Vancouver, BC May 7-12, 2011. A thin outline around the text would allow the date and location to stand out more. Lastly, the overall flow of the logo is nice but I do not think the visual relationship between “20” and “11” is clear. The yellow “11” seems somewhat out of place, and it might be easier to tell that it's intended to be part of 2011 with different placement.”

It is interesting that two workers in the “strength” and “weakness” steps copied content directly from the conference website [Conference] and submitted it as the statement of strength and weakness. The voting scheme in the later step filtered these statements successfully.

Our approach cost $1.41 US dollars and recruited 32 workers to accomplish the critique. It took about 7 hours (1.5 hours for the “description” step, 3 hours for the “strengths” and “weakness” steps, and 2.5 hours for the “summarize” step) to receive the critique. The completion time can be reduced effectively by increasing payment of a task [Heer and Bostock 2010].

The presented critique systematically analyzes the logo design and offers constructive suggestions for change. Compared with the critiques obtained from the online community (see Figure 1), the quality of the critique is significantly improved. We also tested the approach on several visual designs such as posters and the results suggest that the framework is robust and reliable.
3.3 What Was Learned

My preliminary work has yielded lessons for the development of crowd-based systems to support creative activities and resulted in publications in ACM Conference on Human Factors in Computing Systems and ACM Conference on Computer Supported Cooperative Work [Xu and Bailey 2011; Xu and Bailey 2012; Xu and Bailey 2012; Xu et al. 2012; Xu et al. 2013].

- Feedback from online communities is currently ineffective. The results of my first preliminary work revealed salient gaps between users’ expectations and their actual experiences in terms of the quality, quantity and timeliness of feedback [Xu and Bailey 2012]. A majority of online feedback does not go beyond spontaneous judgments such as “I like it”, which offers little value to the user. Unlike professional critics, many users in online feedback communities have insufficient skills, limited time, or little incentive, and thus their individual efforts may not be able to generate a complete critique.

- A paid non-expert crowd has the potential to generate comprehensive design feedback that goes beyond simplistic judgments and matches what designers want, but a feedback task need to be decomposed into micro-tasks doable by people without expertise and consistent with the online labor market.

- The current visual representation of online design feedback is in narrative form. In the next Chapters, I will also explore visual representations of feedback to help designers better understand the received feedback from the crowd.
Chapter 4

Formative Study of Desired Feedback Types

As discussed in Chapter 3, an emerging option for receiving feedback on designs is harnessing a crowd of non-experts. Researchers have recently enabled users to pose design alternatives to a crowd to collect their preferences [Dow, Gerber and Wong 2013]. The advantage is that crowd workers offer a potentially diverse audience, but the drawback is that the workers do not typically have expertise in visual design. To enable crowds to provide additional feedback on visual designs (beyond preferences), it is critical to know what types of feedback users would desire from non-experts and explore how that feedback could be generated by a system and effectively aggregated and presented to users.

Non-expert feedback has long served an important role in the user-centered design process [Blomberg and Henderson 1990]. For example, to better design a computer-based system, designers often present their prototypes to a target audience and collect individuals’ feedback on the prototype. Researchers in service marketing [Johnston 1995; Parasuraman et al. 1985] have demonstrated that customers’ feedback is critical to evaluate the quality of service provided by companies and they have developed desired feedback types, such as the reliability of the service and the courtesy of the service. However, these types cannot be directly applied to visual design domain.

In this Chapter, I will describe a formative study to identify what feedback is desired by users from a crowd of non-experts. These identified feedback types can also help to inform the design of Voyant.

4.1 Methodology

In the formative study, I conducted interviews with twelve participants (six female). Eight were experienced graphic designers (experience ranged from 1 to 8 years; median was 3.5) while the other four occasionally create visual designs (e.g. data graphics, posters for
Inclusion of these two user categories was intended to tap diverse perspectives on what type of feedback is desired during the design process. Interviews lasted about one hour and were conducted face-to-face in the user’s workspace. Participants received $15.

The interviews were semi-structured. The participant was first asked to describe a recent or ongoing design project that served as an initial warm-up and offered context for the discussion. Twelve prepared questions were asked probing the importance of feedback in the design process, what methods are used to receive feedback and from whom, what makes feedback effective, and what type of feedback might be desired from a non-expert crowd (posed as an audience without design knowledge). Additional questions pursued interesting points raised by the participant. Table 4-1 shows the topics covered and a sample of questions in each one. Following qualitative analysis methods [Strauss 1987], the interview data was

<table>
<thead>
<tr>
<th>Characterizing Design Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Could you describe a recent or ongoing design project for which you received feedback? How did this feedback influence the design or your approach to the design?</td>
</tr>
<tr>
<td>How often do you receive feedback for a design project? How would you characterize effective and ineffective feedback?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Types of Feedback Desired</th>
</tr>
</thead>
<tbody>
<tr>
<td>What type of feedback is most desirable during your process?</td>
</tr>
<tr>
<td>How do you usually receive that feedback (informal discussion, formal critique, online community)? What do you see as the key strengths and weaknesses of these methods?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Desirability of Feedback from Non-Experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imagine you could interact with a crowd of non-experts during the design process. What questions would you want to ask the crowd? What type of feedback would you want to receive?</td>
</tr>
</tbody>
</table>

Table 4.1 The main topics and a sample of questions asked within each topic for the semi-structured interviews.

conferences, or personal Web sites) as part of their work or interests but had no formal training in design. Inclusion of these two user categories was intended to tap diverse perspectives on what type of feedback is desired during the design process. Interviews lasted about one hour and were conducted face-to-face in the user’s workspace. Participants received $15.
coded using a bottom-up approach. The data was segmented into logical units and a first pass was made to assign categories to the units. Subsequent passes were then performed to organize the categories into broader themes.

4.2 Importance of Receiving Feedback

I had three main findings from the interviews.

First, all of the participants emphasized the importance of receiving feedback during the design process. Echoing prior work [Cross 1982; Schön 1992; Tohidi, Buxton, Baeccker and Sellen 2006], participants reported that feedback helps them better understand what is (not) being communicated in the design, inform design choices, and remove creative blocks. One participant stated:

“When I get feedback, I start thinking about things in a different way, or I start realizing something that I was overlooking because I spent so much time on the project ... It helps me come to a final thing.” [PU1]

4.3 Experiences of Receiving Feedback via Existing Methods

I found that participants may struggle to receive the feedback desired using existing methods.

Feedback from Instructors in Design Studio

Several participants who were taking academic course in design education have the access to feedback from their instructors; however, instructors’ feedback may increase participants’ feelings of accountability in terms of acting on feedback. Thus, participants feel obligated to correct their designs based on the received feedback. One participant explained:

“There is something about authority figure. If you have a teacher critiques your work, there is a lot of more weight to it. Because you want or you feel you need to be doing it. If a critique from a peer, you don’t necessary feel the pressure as much to change your work.” [PU10].
Moreover, feedback from others knowledgeable about design was deemed useful but not typically available outside of scheduled critiques or studio work, or if one’s peer network does not include designers.

“There is no easy way to access feedback... I don’t have many friends who are designers. For graphic design stuff, I haven’t received extensive feedback... I usually just pick my favorite layout and color” [PU4].

Feedback from Peers

Similar to results reported in [Tohidi, Buxton, Baecker and Sellen 2006], participants reported that feedback received from peers is often overly optimistic and lacks clear direction for improvement:

“They [peers] know what went into it and how hard I worked. Maybe they don’t like my project but they don’t want to hurt my feelings. So those will be sugar-coated” [PU2].

Participants who leveraged online communities such as Dribbble complained the feedback received was too little and lacked depth (echoing findings in [Willett, Heer and Agrawala 2012; Xu and Bailey 2012]). As one participant stated:

“I post my work and I will get feedback occasionally. But it’s usually very shallow and very positive. ‘that looks nice or good work’. It is not very often that you get very true feedback on this site. It’s also a follower-based community. So I have like 50 followers and some people have thousands. So they get a lot of feedback. Since I have so few, I don’t get a lot of feedback.” [PU3].

“As here (online), these people don’t know me and they don’t care. They will just ‘I don’t like it’ or ‘like it’.” [PU12]

Requesting feedback is therefore considered a limited resource that must be used judiciously.
4.4 Feedback from a Large Audience of Non-experts

In the study, all of the participants expressed that feedback from a non-expert crowd would be useful for comparing the perceptions of the crowd to their own expectations. For example, PU5 stated:

“We are trained designers to see certain things in certain ways. But if I can get non-trained people to understand my concept. Really see it. That means I did something really well ... There are a lot of people that have designed artifacts that can be only understood by designers. And that’s a problem. Because the general public should be able to grasp your concept and understand it.”.

Other participants also made similar arguments:

“I’d like to get feedback from people outside of the design world. Because it’s so easy to get caught inside the design bubble. A lot of times, if you are surrounded by designers, you only design for designers.” [PU3];

“It’s so important to always consider a design not from a design standpoint... A majority of public would not get the design concept. Just because it’s a really cool font but it’s not mean anything to them.” [PU9];

“I can create something that it is really cool. But if my parents don’t know how to use it. It’s probably not quite effective design.” [PU11];

In addition, participants reported that a crowd of non-expert feedback is not only useful in the visual design domain but also in other design domains such industry design and interaction design:

“For industry design or user-focused design, audience (non-experts) feedback is really important. Whether is it usable? If it is not, then obviously I’m doing something wrong.” [PU2];

“This would be super helpful for interaction design. Because what if you are trying to design a new layout that has not clicking whatsoever and it’s based entirely on scrolling.
It would be interesting to see how an audience that is just trying to get the information they want from a website would use the layout.” [PU7].

4.5 Types of Feedback Desired from Non-experts

From the responses, I identified four types of design feedback that would be desirable from non-experts.

Impressions

One type of feedback frequently mentioned by participants is capturing first impressions when the design is viewed:

“I want their first instinct response. I think that’s the most important response. Because what you initially see is what change you entire bias of it. If you see a poster, you instinctually think it’s boring you are not gonna put any more time to think about it. I think it’s very interesting to understand how people first interpret things.”[PU6];

“Since I would not be looking for specifically designer critique, I honestly just want them to say the first things that come to their mind. Like if someone says that logo looks like an elephant and it’s not supposed to. That’s something that could be really helpful to know. Or there are some logos that looks like something super dirty but maybe you just don’t notice as a designer. Someone could point that out. I just want to know some raw and unfiltered feedback. I don’t expect them to say ‘that typefaces really sound’...” [PU8];

“What are people’s initial impressions of it?” [PU4];

“What’s your reaction to this or what does this make you think of? ... So I can see if that’s the reaction that I want them to perceive” [PU7].

Communicative Goals

A second type of feedback is whether the crowd understands the communicative goals of the design:
“A group of non-designers would be very useful in testing concepts. Getting feedback on how well your concepts reads” [PU5];

“Does my design convey the right message to the audience?” [PU8];

“Here is my goal. How is this coming across to you? I want to see how overall it perceived.” [PU9];

“How well do my concepts read?” [PU10].

First Notice

A third type of feedback relates to assessing the visual hierarchy of a design:

“What is the first thing people see in a design?” [PU1];

“Does it stand out? What do they first read? What do they first see? What do they think they are gonna see?” [PU6].

Design Guidelines

A fourth type of feedback tacitly mentioned is assessing a design relative to design guidelines, e.g. contrast, proximity, and alignment [Williams 2008]. This feedback was inferred from many responses expressing the need for technical insights about a design and clear direction for how to improve it:

“I want highly technical feedback and constructive feedback...what’s wrong and how to fix it?” [PU11];

“Do the colors clash or complement each other? Are there any distractions?” [PU4].

4.6 Discussion

The outcome of the formative study is a descriptive taxonomy of the types of feedback desired from non-experts. This contribution is significant because my results point to feedback (e.g. interpretation of the user’s goals) that is not directly provided in prior research or existing systems. The results can help system designers to better understand
how to design a system that can generate the desired feedback on visual designs using non-experts.

The resulting taxonomy is not necessarily conclusive. This taxonomy provides a starting point for illustrating the desired feedback from non-experts in a more structured way, and the desired feedback has the potential to be more systematically generated. In the next Chapter, I will describe how the desired feedback types can be generated using crowdsourcing techniques and how the taxonomy of the feedback types can be iteratively improved by incorporating additional users’ inputs.
In Chapter 4, I contributed results from interviews identifying several types of feedback on a visual design that are desired from non-experts. The desired feedback types include the order in which elements are noticed, impressions formed when the design is first viewed, and interpretation of the design relative to guidelines in the domain and the user’s stated goals.

Based on findings from the Chapter 4 and prior work [Xu and Bailey 2011; Xu and Bailey 2012], in this chapter, I explore techniques for designing a system that can generate the desired feedback types using online crowds. Specifically, I designed and implemented a novel Web application called Voyant that generates design feedback desired from non-experts. The system will include two major components: feedback generation and feedback interaction. The system accepts a visual design as input, decomposes the feedback generation process into sub-tasks suitable for non-experts, aggregates the results, and presents them to the user. For the non-expert crowd, the system currently uses paid workers from an online labor market.

5.1 Design Goals

I designed Voyant according to the following goals.

Configure the Audience

Receiving feedback from a desired audience is a prerequisite for users to receive the desired feedback on their designs. To help users configure the audience, the system allows a user to be able to specify the demographics of the desired audience. The system must then ensure that the crowd workers who express interest in performing the feedback tasks match the demographics specified by the user. To accomplish this goal, the system generates a demographic survey that workers must complete prior to performing any task. Only if the
worker’s responses match the user’s specification is the worker allowed to progress to the feedback tasks.

**Allow Annotation of Design Elements**

Non-experts do not speak design language and they may not be able to precisely use free-form text to associate their perceptions of a design with specific visual facts in it. As a result, users may have difficulties in understanding the feedback, especially when the feedback is from a large number of non-experts. The system purposely reduces this reliance by allowing crowds to annotate locations in a design to associate with their perceptions. This structured data also enables multiple views in the system for users to understand crowd’s perceptions from different perspectives.

**Support Efficient Feedback Exploration**

Unlike unstructured, free-form comments generated from online communities, the feedback data generated by my system is structured and multidimensional. The system should utilize the unique feature of the feedback data and allow users to interactively explore the feedback data from different perspectives. For example, the interface should allow users to explore associations between the crowd’s perception of a design and the graphic annotations within it. Research in information visualization shows that multiple coordinated views enable users to rapidly explore complex data and discover associations, and empirical evidence suggests that the set of coordinated views can better support exploratory data analysis than the sum of the individual views [Baldonado, Woodruff and Kuchinsky 2000]. Following past work, I design coordinated views to help users explore the feedback. Specifically, for each feedback type, the system will provide coordinated views that consist of a visual overview of the crowd’s perceptions and graphic annotations overlaid on the design.

**Structure the Feedback Generation Process**

Inspired by theoretical frameworks of feedback in design education [Feldman 1971; Feldman 1973; Trumbo 1997], the system decomposes feedback generation into a *description* and *interpretation* phase. The main idea is to convert the feedback process into a series of sub-
Figure 5.1. The interface for configuring the feedback generation process. From this interface, Alice uploads her poster image, configures the desired audience (M/F, USA, 18-35) and selects which feedback is desired. For example, after selecting Goals, Alice enters her communicative goals for the poster.

5.2 Usage Scenario

To demonstrate the utility of the system, I first describe the design feedback that can be generated, interactions for exploring the feedback, and how the feedback benefits the user. Consider a real scenario adopted from the evaluation. Alice, a novice designer, is designing a

... (rest of the text continues)
poster for a dessert station as part of a design project (Figure 5-1). The goal of the design is to attract attention and create excitement about trying the new service in a restaurant. Alice wants to assess how the poster is perceived by her intended audience and how these perceptions align with her expectations. Figures 5-1, 3, and 4 illustrate the system in context of this scenario.

To initiate feedback on a design, the user navigates to the main page of the Web application (www.crowdfeedback.me). From the page, the user uploads a visual design (e.g. poster, logo, or Web page) as an image, selects the feedback desired, and configures any related parameters. Five types of feedback can be selected or configured:

- **Elements** are the individual elements that are visible or otherwise perceived (“seen”) in the design including colors, shapes, objects, and activities.
- **First Notice** refers to the visual order in which elements are first noticed in the design.
- **Impressions** are the perceptions formed in one’s mind upon first viewing the design.
- **Goals** refer to how well the design is perceived to meet its communicative goals. If selected the user is prompted to briefly state each of her goals for the design.
- **Guidelines** refer to how well the design is perceived to meet known guidelines in the domain. The set includes proximity, alignment, repetition, and contrast [Williams 2008].

In early prototypes, the Elements were produced as part of the workflow for generating First Notice. However, user testing revealed that knowing the set of elements seen in the design was useful on its own and was therefore extracted into its own type of feedback. The user may also configure the geography, gender, and age of the audience.
Figure 5.2. The user interface offers coordinated views for each type of feedback generated on Alice’s poster. In the *Elements* feedback, she reviews the list of elements seen by the crowd.

Figure 5.3. *First Notice* feedback. The user views which elements and corresponding regions of the design are first noticed and creates a filter to determine which elements relate to the center markers.
Once the feedback is configured, the system instantiates the crowd workflows, decomposes the complex process of feedback generation into micro-tasks, and submits them to an online labor market (Mechanical Turk). Analogous to [Willett, Heer and Agrawala 2012], the tasks are designed such that they can be performed by crowd workers with little or no design knowledge. The individual task outcomes are then aggregated and presented to the user. Each type of feedback typically requires a few hours to generate and costs a few US dollars. To ensure that the crowd workers match the audience specified by the user, prospective workers first complete a demographic survey and only those workers whose responses match the user’s specification are allowed to progress to the feedback tasks.

For each feedback type, Voyant provides coordinated views that consist of a visual overview of the crowd’s perceptions (e.g. in the form of a list, word cloud, or bar chart) and visual markers overlaid on the design (see Figures 5-2, 5-3, 5-4, 5-5, 5-6 and 5-7). Though coordinated views is an established visualization technique for exploratory analysis [Baldonado, Woodruff and Kuchinsky 2000], we apply this technique to enable analysis of the association between the perceptions of a design and the visual elements within it. Once generated, the feedback is presented as a tabbed interface where each tab contains one type of feedback.

For Elements, the user receives a list of the elements “seen” by the crowd. To help the user parse the elements, they are grouped into four categories including object, color, shape, and activity, which were derived from [Feldman 1981]. The set of categories could be easily modified in future work. Each element can be selected to view the regions of the design associated with it. Marked regions appear as visual markers (circles) overlaid on the design. For example, in Figure 5-2, selecting the element “straw” shows the corresponding visual markers placed by the workers. The benefit is the user can compare what was “seen” to what was expected. For example, in Figure 5-2, some crowd workers perceived the yellow background shape in the bottom right as the “sun.” The user can assess whether this perception aligns with her expectation and, if not, can use this discrepancy to seed the next iteration on the design.
Figure 5.4. Impressions feedback. A word cloud represents the impressions and their frequency that the crowd had upon first viewing the design.

Figure 5.5. Goals feedback. A set of interactive charts show the distributions of the crowd ratings for each communicative goal.
The second tab *First Notice* shows the visual order in which the elements are seen. The list can be ordered globally based on the number of workers who reported seeing the element first or ordered separately in each category. Assessing the visual hierarchy is an important aspect of visual design [Kress and Leeuwen 1996]. For example, from the feedback shown in Figure 5-4, the user learns that the glass with the red-colored content is the element most frequently seen first as indicated by the concentration of visual markers and the corresponding elements “red,” “pink,” “drink,” “glass,” and “beverage.” Interestingly, “fresh smoothie” attracts far less attention despite its prominent size and location.

The *Impressions* tab shows a word cloud of the crowd’s first impressions of the design (Figure 5-5). A word cloud is used as it offers a compact representation and efficiently communicates the more and less frequent impressions. The word size is proportional to the number of workers who shared the impression. If the user selects a word, the system shows the visual markers on the design corresponding to that impression. Conversely, the user can select an individual or group of markers on the design image to view the corresponding impressions in the word cloud and read the explanations offered by the workers. For example, from Figure 5-5, one of the most common perceptions was “retro” which was associated with the title and the red and brown drink glasses in the poster. Though not shown in the figure, the user could also explore the markers placed on these elements to identify other words associated, e.g., the glasses were associated with “colorful”, “fresh”, and “juice”.

The two remaining tabs for *Guidelines* and *Goals* provide similar forms of feedback and interaction. For each of these, Voyant presents an interactive bar chart summarizing the ratings from the crowd. The ratings were made on a seven-point scale from strongly disagree (-3) to strongly agree (+3) with whether the design met the given guideline or goal. Also, each worker could place a marker to indicate the region of the design associated with the rating and enter rationale. The user can select subsets of the guidelines or goals to inspect and compare the results (Figures 5-5, 5-6 and 5-7).

In addition, the user can select a subset of the bars in a chart to inspect only those markers and explanations associated with the selected ratings. For example, in Figure 5-6, the user selects all the non-positive ratings for *contrast* in the chart and is then able to review the corresponding visual markers and explanations. Likewise, as shown in Figure 5-7, the user
can select a set of visual markers on the design to view the corresponding ratings and explanations.

Figure 5.6. *Guidelines* feedback. A similar set of charts shows the distributions of the ratings for each design guideline. The user selects the non-positive ratings to inspect the corresponding explanations and visual markers.

Figure 5.7. *Guidelines* filtered. The user further explores the feedback by selecting a subset of the visual markers (bottom center), which filters the ratings shown in the chart and the explanations.
From the ratings of the goals shown in Figure 5-5, the user can have increased confidence that the goals are being reasonably communicated given the positive ratings. But, from the ratings of Guidelines in Figure 5-6, the user learns that the design could use improved contrast. Inspecting the explanations reveals that many crowd workers felt the color and style of the “fresh smoothie” text did not adequately stand out relative to the background and other colors. This and other feedback can seed additional design iterations.

5.3 Discussion

For each type of feedback generated by Voyant, the system provides multiple coordinated views that consist of a visual overview (in the form of a bar chart, word cloud, or list) and visual markers overlaid on the design that couple the overview to specific locations on the design. This unique feature of our system helps users analyze the association between the crowd’s perception of a design and the visual elements within it. For example, the site http://fivesecondtest.com produces a word cloud similar to the word cloud produced in our system for the Impressions feedback. However, unique to our system, the user can select specific impressions in the word cloud to identify the spatial locations on the design associated with those impressions. Conversely, the user can select any visual markers on the design to view the associated impressions in the word cloud. The coordinated views allow the user to explore the feedback spatially (within the design) or from the overview and such views are provided for each type of feedback generated in our system.

The current system only allows users to configure the geography, gender, and age of the audience. However, other configurations such as occupation, habits and social background can be helpful for users to specify the audience of their designs. These configurations can be integrated into the system in the future.

The usage scenarios described in this chapter reveal interesting details into how users can interact with crowd feedback using multiple coordinated views provided by Voyant. Figures 5-2, 5-3, 5-4, 5-5, 5-6 and 5-7 reflect the end-state of each scenario, but many of the earlier steps in the scenario are visible as well. Additional details about how users can benefit from the feedback and the interactions provided by Voyant will be presented in next chapters.
Chapter 6
Crowd Workflows for Feedback Generation

The Chapter 5 describes the user experience of Voyant, a novel system giving users access to a non-expert crowd to receive feedback on their designs from a simulated audience. Voyant generates the elements seen in a design, the order in which elements are noticed, impressions formed when the design is first viewed, and interpretation of the design relative to guidelines in the domain and the user’s stated goals. The coordinated views in Voyant are designed to help user analyze relations between the crowd’s perception of a design and the visual elements within it.

In this chapter, I will illustrate how the feedback types desired by users are generated using a crowd of non-experts. Each feedback type is generated by a customized workflow to meet the constraints of what users need. Also, since the system uses a non-expert crowd to construct the desired feedback, it is necessary to provide at least some scaffolding (e.g. by designing focused tasks consistent with the norms of a micro-task labor market) to account for workers’ abilities and construct workflows that effectively aggregate their individual efforts.

The crowd workflow I developed was inspired by how novices are taught to perform critiques of visual designs taught in design education [Feldman 1981] and the design of the worker tasks followed the recommendations provided by [Willett, Heer and Agrawala 2012]. Specifically, Voyant decomposes feedback generation into a description and interpretation phase. In each phase, the tasks focus worker attention on specific aspects of a design rather than soliciting holistic evaluations to improve outcomes.

6.1 Description
The purpose of the first phase is to enumerate what can be “seen” in a design such as objects, colors, shapes, and activities. The inclusion of this step was motivated by how novices are taught to critique visual designs by first describing what elements are seen [Feldman 1981].
Once identified, one can consider how the elements relate to each other, the goals of the designer, or guidelines in the domain.

To generate elements for a design, micro-tasks are created and submitted to the online labor market. The worker task screen shows the design and prompts the worker to enter at least two elements seen in the design relating to a given category (color, object, shape, or activity). These categories can be expanded by the user. The category prompts provide a means for focusing worker attention on the task (see Figure 6-1a). After a worker lists two or more elements, s/he is asked to mark regions of the design corresponding to each element listed, if applicable (Figure 6-1b). The markers provide an efficient mapping from what is described to what is seen. Five tasks are created for each category by default. However, if elements collected for a category are all unique, additional tasks are created for that category until a new element repeats an earlier one (a proxy for saturation). The repeated element is removed and the final set is piped into the second phase - interpretation.

6.2 Interpretation

In the second phase, the design is interpreted relative to the visual hierarchy, first impressions, design guidelines, and any communicative goals entered by the end user.

6.2.1 First Notice Workflow

The workflow for First Notice creates task screens showing the design and all the elements listed during Description. The ordering of the elements on the screen is randomized to avoid bias (Figure 6-2). The worker is prompted to select which element s/he notices first in the design. The worker is then instructed to mark on the design (by placing and sizing a rectangle) where the selected element is located. From experimentation, the number of workers currently recruited is twice the number of elements listed, up to a limit of thirty. The number of times each element is selected is returned and presented to the user (Figure 5-3, right panel) along with the corresponding visual markers (Figure 5-3, left panel).
Look at the design carefully.

Name **at least two objects** you see in the design. Use a comma(,) to separate the objects. If an object is not in the design, your result could be rejected.

```
straw, glass
```

(a)

Draw a box on the following design to indicate: **straw**

If your drawn box is not related to the element you described, your result could be rejected.

(b)

Figure 6.1. The workflows and related task screens for generating *Elements* feedback in Voyant. (a) the worker enters at least two elements seen or perceived in the design and then (b) sizes and locates a rectangle to locate each element.
Look at the design and the word cloud carefully.

Please select the word from the list below that best describes the **first element** you see in the design.

- red
- white
- cone
- sun
- glass
- sunbathing
- brown
- sipping
- yellow
- hourglass
- beverages
- smoothie
- green
- rectangle
- straw
- blue

Figure 6.2. The task screen for generating *First Notice* feedback in Voyant. A worker selects which element is first noticed in the design.

### 6.2.2 Impressions Workflow

The *Impressions* workflow generates task screens showing the design and prompts the worker to enter one or more words that come to mind when first viewing it (Figure 6-3a). For each word entered, the same worker is prompted to mark the design (place and size a rectangle) to indicate the region related to the impression and enter rationale (Figure 6-3b). As an aid, the system inserts up to three elements from the *Description* phase whose markers are nearest to the one just entered. This workflow ends once twenty tasks have been completed or thirty unique impressions are collected.

From the impression words collected, the system extracts the unique words and their frequency. Additional workers are then recruited to view the design, select which words from the existing set best captures their impression (Figure 6-4), and enter rationale. This latter step is performed to further elicit the frequency of impressions from the crowd.
Look at the design carefully.

What are the first words that come into your mind to capture your impression about the design? Use a comma(,) to separate the words if you enter more than one.

(a)

retro

submit

Draw a box on the following design to indicate: retro

Describe the elements in the part of the design you just highlighted, such as smoothie, pink, and drinking. Explain why your impression relates to these elements (at least 15 words):

Drinks, hot pink, lime green. It looks like an ad you might have found in the eighties especially the yellow "flash" around the drinks. The font doesn't seem especially retro though

(b)

Figure 6.3. The workflows and related task screens for generating Impressions feedback in Voyant. (a) a worker enters the first word(s) that come to mind after viewing the design. For each impression entered, (b) the worker locates the most related area of the design and enters rationale.
Look at the design and the word cloud carefully.

Please select the word from the list below that best captures your **first impression** when viewing the design.

<table>
<thead>
<tr>
<th>clean</th>
<th>tropical</th>
<th>delicious</th>
<th>drink</th>
<th>chocolate</th>
<th>holiday</th>
<th>fresh</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple juice</td>
<td>old fashioned</td>
<td>colorful</td>
<td>strawberry</td>
<td>retro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>energy drink</td>
<td>fruity</td>
<td>tasty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.4. The task screen for generating Impressions feedback in Voyant. Additional workers vote on the collected impressions and enter rationale.

Read the design guideline and compare negative and positive examples.

**Guideline:** A design should closely group elements that have relationships. The design should use space to clearly define groups of elements.

**Negative example:** None of the elements connect with any other element in the design.

**Positive example:** This design groups related elements together.

Figure 6.5. For **Guidelines** feedback, a worker learns about the guideline Proximity.
Evaluate the design by the following guideline carefully.

**Guideline:** A design should closely group elements that have relationships. The design should use space to clearly define groups of elements.

**Statement:** The design meets this guideline.

strongly disagree moderately disagree slightly disagree neutral slightly agree moderately agree strongly agree

Submit

(a)

Draw a box on the design to indicate which part of the design supports your rating.

Describe the elements in the part of design you just highlighted, such as **beverages, pink, and glass**. Explain why these elements in this part of the design meet the guideline (at least 15 words):

The glasses are together and in the same direction. They have relationships and are closely grouped.

Submit

(b)

Figure 6.6. Related task screens for generating Guidelines feedback. (a) After a worker learn a design guideline (see Figure 6-5), he rates the design on the guideline, and (b) locates the area of the design that best supports the rating and enters rationale. The workflow for Goals is the same as that for Guidelines.
For the Guidelines workflow, a series of tasks are generated where workers rate how well the design adheres to the set of guidelines available in the system. A worker rates his or her perception of how well the design adheres to a guideline on a 7-point scale (Figure 6-6a). The worker is then prompted to mark a region on the design and describe how the content in that region relates to the rating (Figure 6-6b). This task is repeated several times per guideline, and then repeated for all guidelines selected by the user during the configuration phase. Because workers are not assumed to have knowledge of visual design, the first part of each task screen explains the guideline to be rated in lay language and shows both a positive and negative example (Figure 6-5).

6.2.4 Communicative Goals Workflow

The workflow for Goals is similar to Guidelines. A series of tasks are created for each goal entered by the user during the configuration phase. A worker views the design and one of the goals and rates how well the design is perceived to meet the goal. As with Guidelines, the
worker then marks a region on the design that best relates to the rating and enters rationale. This task is repeated several times per goal.

The micro-tasks generated by the system are submitted to Amazon.com’s Mechanical Turk, a popular online labor market. Table 6-1 shows the default number of workers and total payment made for each feedback type. These values were determined through experimentation and could be easily adjusted or made configurable by the user. Only workers with >= 95% task approval rate are considered for our tasks.

6.3 Discussion

The five types of perception-oriented feedback generated by my system are not directly provided by existing commercial systems. For example, existing crowd-based usability sites [Feedbackarmy; Fivesecondtest; Usabilla] mainly help designers conduct task-oriented tests and surveys of their Web designs. When using these sites, the user must conceptualize and phrase the questions that will be posed to the crowd to collect worker responses. It is possible, however, that the open question format of these sites could be leveraged to approximate the questions targeted by my system (e.g. the user could write questions such as “what elements do you see in the design?” or “what is your first impression of the design”). The problem is that this approach assumes the user knows which questions to ask and is able to phrase them in a way that yields meaningful responses. In contrast, since Voyant generates specific types of feedback, the user does not have to conceptualize and phrase the questions, the system can tailor the visual representations for each feedback type, and we can design the crowd workflows and worker tasks to accommodate what is doable by non-experts and methods for aggregating the individual task outcomes.

The current system did not attempt to optimize the workflows to reduce either the time or cost of generating the feedback. Also, this feedback might be improved with further scaffolding for the crowd workers or integration of innovative workflows.

The current system uses the crowdsourcing service provided by Amazon.com’s Mechanical Turk (MTurk) to receive workers and all the micro-tasks generated by the system will be submitted to MTurk. MTurk can provide flexible and abundant labor force, and an API for
easily creating and removing micro-tasks. However, the system could use other crowds such as the employees of a large organization, members of an online community, or students in a large design course.
Chapter 7

System Implementation

The Chapter 5 and 6 describe the design of a system (Voyant) for generating structured feedback on visual designs from non-experts. I mainly emphasize user experience of the system and the crowd workflows for generating that feedback.

In this chapter, I will focus on the technical implementation of the system. The functionality described in the Chapter 5 and 6 has been fully implemented. The system infrastructure consists of two entities: feedback generation and feedback interaction (see Figure 7-1). Although there are many ways to implement the system, I built the system as a Web application. Because users may use various kinds of graphic software when they perform visual design work. Web applications do not rely on client side software beyond the browser, do not require remote installation, and provide an interface the user is already familiar with.

The current system is built as a Web application and can be accessed from most popular Web

![Diagram](image)

Figure 7.1. System infrastructure overview. Voyant provides users on-demand access to design feedback from groups of non-experts. The user formulates her feedback request on a design through a web-based interface. The system instantiates the workflow to generate the configured feedback by recruiting workers to perform the set of Describe and Interpret micro-tasks. The system aggregates the task outcomes and allows users to explore the crowd based feedback through visual analytics.
browsers.

I adopted an iterative design process to develop the proposed system and each aspect of the system went through many changes. I started by proposing several conceptual designs with low-fidelity prototypes. I collected users’ feedback and improve the system at each iteration. For example, the system initially only provides four design feedback types. The Elements were produced as part of the workflow for generating First Notice. However, user testing revealed that knowing the set of elements seen in the design was useful on its own and was therefore extracted into its own type of feedback. Thus, the current system provides five design feedback types.

7.1 User Interface Implementation

The user interface was built using JavaScript and D3.js [Bostock et al. 2011]. The core part of the interface in Voyant is multiple coordinated views, which is designed for users interact with crowd feedback. In order to support browsing multiple coordinated views of the feedback data, the interface should provide flexible navigation through the multi resolution data hierarchy and gives the user full control by providing key features such as brushing and linking, and focus and context. There are a few notable visualization libraries that can be used to build the envisioned user interface with multiple coordinated views rapidly.

Processing.js

Processing.js, which complements to the Processing project, is a library for immediate-mode graphics that uses HTML5 Canvas. It can dynamically generate static and interactive graphs and charts in the user's web browser. Processing.js can be used to create powerful animation in 2D or 3D. However, it does not generate vector graphics or HTML elements.

Raphael.js

Raphael.js is a JavaScript library, which is used to generate vector graphics in the browser. It can be used for the low-level Scalable Vector Graphics (SVG) rendering operations such as lines and points. Raphael.js also provides an easy way for system developers to manipulate graphics (e.g. scaling). Raphael provides many prepackaged layouts and developers might be
limited to a set of methods. In addition, although Raphael.js has good documentation to use for the implementation, it is not updated in a timely manner.

**D3.js**

D3.js is also an open source low-level library for generating SVG-based charts. D3.js provides a very flexible way to specify elements as HTML or SVG directly. This can allow developers to create new customized visualizations from scratch instead of having to select from a predefined features. More importantly, D3 focuses on transformations between scenes, and thus it helps developers easily generate animation, interaction, complex and dynamic visualizations. I decide to use the D3.js library to implement multiple coordinated views for the following reasons:

- **Flexibility.** D3.js can help me easily create customized and complex visualizations. This is important because multiple coordinated views for feedback requires various new layouts and interactions, which do not resemble common charts predefined in other libraries.

- **Compatibility.** D3.js is based on the current web standards and is compatible across browsers. More importantly, D3.js provides an efficient way to bind data to elements.

- **Well-established Documentation.** D3.js has been widely adopted by developers. The library is frequently updated, and bugs or issues in the library can be fixed in a short time. Also, there are many well-documented examples, which help developers implement complex visualizations.

**7.2 Server**
The backend was constructed using Ruby-on-Rails [Ruby-on-Rails] and MySQL [MySQL]. A MySQL relational database is used as the data store for user, worker, design, feedback features, and event data.

Ruby on Rails is an open-source framework for web applications. I adopted Ruby on Rails for the following reasons:

**Simplicity**

Rails have automated many things in the framework. So developers can just mainly focus entirely on solving the problem. Ruby on Rails is also a simple development environment that provides transparent support for AJAX and built-in Object-Relational Mapping database support.

**Maintainability**

Ruby on Rails is well supported and has an active community. The Rails community provides various plugins as Ruby gems, which can be easily installed to developers’ own projects. For example, Voyant needs to have a fairly standard user authentication / permission system. Devise is a gem which can be simply added to Voyant for authentication and user management.

Figure 7.2. One of the initial prototypes of Voyant created after initial discussion with my advisor and teammates.
MVC Architecture

Ruby on Rails framework is based on the Model View Controller (MVC) architecture, which provides a clear separation of presentation from data. Model maintains the relationship between database and object. Model provides an interface and binding between the tables in a relational database (see Figure 7-3) and the Ruby program code that manipulates database records. View is the presentation layer for a Web application. For example, the source code of the front end of Voyant is mainly in the “View” folder (see Figure 7-4). Controller is the connection between Model and View. For example, Controller processes the queries from Model for specific data, and then organizes the data into a form that fits the needs of View.

Voyant is now available for public (www.crowdfeedback.me). Releasing the system for the public by itself is a challenging problem which requires a lot of engineering efforts. GIT was used as the backend repository format to track changes and find the origin of each line of code. Figure 7-5 shows that each aspect of the system went through many changes. For example, a design may have a large amount of feedback data and the size of the data file is too large to be loaded from the server. I implemented compression algorithms to reduce the size of data file and the reliability of the system was improved.

<table>
<thead>
<tr>
<th>Lines of code</th>
<th>User Interface</th>
<th>Workflow</th>
<th>Server</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30,000</td>
<td>12,000</td>
<td>40,000</td>
<td>82,000</td>
</tr>
</tbody>
</table>

| Files in master   | 507            |
| Commits           | 656            |
| Branches          | 14             |

Table 7.1. Implementation Statistics of Voyant.
Figure 7.3. Entity relationship diagram is used to visualize the relations among the tables in the database.
Figure 7.4. A tree view for folders in Voyant.
Figure 7.5. A series of commits were made to Voyant. These commits including editing the feedback request page, displaying feedback generation speed, and fixing bugs with IE.
7.3 Discussion

Voyant can be used by both designers and researchers. Designers can use Voyant to receive feedback on their creative work and iterative toward a more effective solution. Voyant can also be used as an experimental platform for research on crowd feedback systems. Using Voyant, it is possible to study the effects of crowd feedback in a longitudinal and large-scale context.

Ruby on Rails is a popular framework for web application. However, I found that it actually has a steep learning curve. Developers need not only to understand the Rails framework, but also to learn a series of languages such as CoffeeScript, SCSS, and HAML. Also, Ruby is worse than Java in terms of scalability and stability.

One lesson for front-end web development is that D3 does work well with Internet Explorer (IE). This library often contains code which causes errors with IE8 and below. However, I found that a significant portion of crowd workers on Amazon Mechanical Turk are using IE. Developers should consider how support older browsers when they implement interfaces for crowd workers.

In the future, the system could provide public APIs for requesting and receiving crowd feedback. Then, we can easily embed crowd feedback into graphics design software such as Adobe Photoshop and Illustrator, and users can get feedback without leaving the graphics software.
Chapter 8

Evaluation

Chapters 5-7 describes the experience, workflows and implementation of the system, Voyant, for generating structured feedback on visual designs from a non-expert crowd. The system can help any user who creates visual designs receive structured, perception-oriented feedback from a selected audience.

The next challenge is to assess the value of the feedback, as well as the impact on users and their designs. To address this challenge, in this chapter, I reported on the outcomes of three studies to evaluate Voyant from three perspectives. First, I carried out a lab study to understand users’ initial perceptions of the system and the feedback. Then, I conducted a field study to understand how Voyant can help users improve their designs in practice. Finally, I performed a content analysis to examine differences between Voyant feedback and free-form feedback.

8.1 Initial Perceptions of Voyant

My first study was designed to understand users’ perceptions of the benefits of the feedback generated, the usability of the visual feedback representations, and interactions in the interface. Formally, this study attempts to address three research questions.

RQ1: What are the users’ perceptions of the Voyant feedback overall? What do they learn from the received feedback?

RQ2: What are the users’ perceptions for different types of the feedback generated by Voyant?

RQ3: How is the usability of the usability of the visual feedback representations, and interactions perceived by users?
8.1.1 Methodology

Eight users (PL13-20) were recruited for the study (4 female) and none had prior knowledge of our system. Four of the participants were practicing graphic designers (2-5 years of experience) who worked on print, Web, and interaction design. The other four created visual designs for personal use such as Web pages and illustrations as part of course or research projects. They had no training in visual design.

I collected one in-progress or recently completed visual design from each participant for which feedback was desired. Figure 8-1 shows the designs collected. For example, PL14 was designing a magazine cover and wanted to visually communicate it is about a film festival. PL17 was designing a Web site showing positive social work in the community and wanted to convey a warm, inviting feeling. Relevant information such as the goals and intended audience was collected for each design. Our system was then configured and used to generate the full set of feedback on each design. The time to generate the feedback on each design ranged from 22 to 70 hours with a median of 40 hours. The cost ranged from $10.70 to $13, with a median of $11. Interviews were then scheduled with the participants. An interview lasted one hour and remuneration was $15.

The interview consisted of a warm-up (15 minutes), system walkthrough (30 minutes), and post discussion / survey (15 minutes). In the warm-up, the participant described the feedback previously received on the design, where it came from, and how useful it was. For the walkthrough, I first introduced the system and then the participant used it to explore each feedback type at their own pace while thinking aloud. I prompted discussion about the helpfulness of the feedback, usability of the interface, and how the system can be used in practice. Afterward, the participant completed a survey. It contained eleven questions stated in a neutral manner regarding the helpfulness of the system and each type of feedback. Responses were given using a 7-point Likert-scale ranging from strongly disagree (1) to strongly agree (7). Table 8-1 summarizes the questions and responses.
Figure 8.1. The designs collected for the evaluation included Web designs, a magazine cover, and an illustration.
Participants reported that the feedback generated by our system was helpful overall and would help them improve their designs (Q1, $\mu=5.8$). For example, participants stated:

<table>
<thead>
<tr>
<th>Question</th>
<th>Avg (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. The feedback would help me improve my design</td>
<td>5.8 (1.3)</td>
</tr>
<tr>
<td>Q2. The “Impressions” feedback was helpful</td>
<td>6.0 (1.2)</td>
</tr>
<tr>
<td>Q3. The “Goals” feedback was helpful</td>
<td>5.4 (1.2)</td>
</tr>
<tr>
<td>Q4. The “First Notice” feedback was helpful</td>
<td>5.1 (1.0)</td>
</tr>
<tr>
<td>Q5. The “Elements” feedback was helpful</td>
<td>4.5 (0.9)</td>
</tr>
<tr>
<td>Q6. The “Guidelines” feedback was helpful</td>
<td>4.1 (1.6)</td>
</tr>
<tr>
<td>Q7. The user interface was easy to learn and use</td>
<td>5.2 (1.6)</td>
</tr>
<tr>
<td>Q8. The overview for each feedback type was helpful</td>
<td>6.2 (0.7)</td>
</tr>
<tr>
<td>Q9. It was easy to inspect details of the feedback</td>
<td>5.5 (1.1)</td>
</tr>
<tr>
<td>Q10. The use of the visual markers was helpful</td>
<td>5.4 (1.4)</td>
</tr>
<tr>
<td>Q11. I would request feedback more often with Voyant</td>
<td>5.1 (1.1)</td>
</tr>
</tbody>
</table>

Table 8.1. Participant ratings for the different types of feedback generated and the overall system.

### 8.1.2 Helpfulness of the Feedback Overall

Participants reported that the feedback generated by our system was helpful overall and would help them improve their designs (Q1, $\mu=5.8$). For example, participants stated:
“There is some misunderstanding of my poster [see Figure 8-2]. Many people said that the portrait is not clear enough to show the Shakespeare’s figure. They think it’s a dog. That’s the most important part I learn from the feedback. I will make new sketches based on the feedback.” [PL18]

“It is useful to see the number of people that have certain issues. I keep getting people’s misunderstanding of the icons of the drink specials [see Figure 8-3]. I definitely get right away that was a major issue because there are so many people that talked about it” [PL20]

“This is very useful feedback and something that I couldn’t get in any other way.” [PL17]

Figure 8.2. The Impressions feedback generated by Voyant on a user’s poster. The user intended it to be perceived as Shakespeare, but was surprised to learn of an unintended interpretation (see “dog” in word cloud). The interface is showing a word cloud representing the impressions formed when the crowd first viewed the design. Selecting an impression (in orange) shows the corresponding explanations and visual markers associating specific regions of the design.
Figure 8.3. Part of the feedback (Goals) generated by Voyant for a Web design collected from PL20. The participant explored the non-positive ratings and learned that the icons were not being perceived the way he intended and needed refinement.

Figure 8.4. One design goal of the 3D rendering is render teapot in Chinese style. The participant learned that this goal was not well supported. After exploring the visual markers and explanations, she decided to add tea cups beside the teapot and add a dragon pattern to the teapot to better communicate Chinese style.
Specifically, the feedback helped users identify problems with their designs. For example, one of PL15’s design goals was to create a rendering of a teapot in Chinese style (see Figure 8-4). However, she learned from the Goals feedback that this goal was not well supported. After exploring the visual markers and explanations linked to the lower ratings for this goal, she gained a better understanding of the problem. To address it, she stated she would add tea cups beside the teapot to reinforce the perception of a teapot rather than a coffeepot and add a dragon pattern to communicate Chinese style. Likewise, PL20 discovered from First Notice that

“Not a lot attention is given to the main navigation bar at the top... We need to do something to highlight this area more.”

Due to the low ratings and the corresponding explanations about contrast in Guidelines, PL16 said that he would now increase the contrast between the foreground and background colors.

A second way the feedback was helpful is that it allowed participants to gain insight into their designs. For example, in the poster created by PL19 (see Figure 5-5), her color selections were intended only to make the poster “colorful.” But from the feedback, she discovered that people were associating different flavors with her color selections (e.g. mango or apple to the yellow-orange smoothie) and that she could manipulate the color to cause people to literally ‘taste’ her design.

A third way the feedback helped is that it allowed participants to consider their designs from different perspectives, especially when the feedback contained conflicting viewpoints. For example, PL13 said

“Some of these positive things are contrasting what negative people said [see Figure 8-5]. I find myself comparing “+3” to “-3” if they are talking about the same thing... I can see why it worked for one person and see why it didn’t work for another person. So it’s easy for me to find a compromise between these two.”

Finally, participants felt the feedback was a useful complement to prior feedback received on their designs:
“I think the quality of the feedback is really good. Because many parts of the feedback are actually the same as the feedback I got from my professors and my classmates.” [PL15]

“It is really interesting to get people’s feedback, especially if it’s coming from non-designers. I’m so used to focusing so much on type. The most surprising thing is that people [non-designers] don’t care much about the type as they care about the imagery.” [PL14]

“I got a lot more critical feedback. That’s very hard to get from peers.” [PL16]

### 8.1.3 Helpfulness of the Different Types of Feedback

The Impressions feedback was reported as the most helpful among the five types of feedback generated (Q2, $\mu=6.0$). From the perspectives of the participants:
“I like this word map [Impressions] and I think this word map is incredibly helpful... Because I can think about what I was trying to achieve with this design and reflect it to the words people are using when describing this. Then I can see if I really did well enough by portraying something in my design.” [PL13]

“Impressions were the most useful because they brought up more unexpected details.... I’m looking for unusual words or data. I can see the less common terms associated with the design.” [PL19]

Participants also agreed that the Goals (Q3, µ=5.4) and First Notice (Q4, µ=5.1) feedback was helpful:

“Goals is helpful for verifying the level achievement of goals that are actionable by the audience. Goals sort of remind me of AB testing if I had multiple iterations of the design and compared the feedback on the goals.” [PL17]

“What you see first is important. This [First Notice] will give me information on how to make things more attractive and memorable in short time frame.” [LP19]

There was moderate agreement the Elements feedback was helpful (Q5, µ=4.5). PL15 felt the feedback was “a good check” on whether design elements are interpreted correctly. He reasoned that

“At a basic level, those things [Elements] are very important to understand. Like if someone says that object looks like an elephant, and it’s not supposed to. That’s something that could be really helpful to know.” [PL15].

Other participants felt the feedback may be too low-level.

The Guidelines feedback received lower ratings, though the average was still on the positive side of neutral (Q6, µ=4.1). Positive reactions included:

“This is useful, especially for me. I’m bad at following common design principles” [PL20]; “because I’m not a professional designer. I might forget about color, positioning, or grouping elements. If they mentioned these, it’s useful for me to keep these in mind.” [PL17].
In terms of less positive reactions, some participants had strong pre-conceptions and “don’t necessarily trust non-experts to give technical feedback” [PL18]. However, after reviewing the Guidelines feedback on their designs, they became less skeptical:

“I like this comment. This line is not parallel to that line. It doesn’t satisfy the alignment design guideline. That’s a good point.” [PL18].

Finally, two participants commented that some explanations about the given ratings were “arbitrary” or “not related” and they had to expend effort to filter these irrelevant explanations.

### 8.1.4 Feedback Representations and Interactions

The coordinated views provided by Voyant enable exploration of the feedback from the perspective of either the overview representation or the markers on the design. Participants reported that this interactive coupling was helpful for analyzing the association between the crowd’s perception of a design and the elements within it (Q7-Q10, μ>=5.2). For example, when exploring Impressions, I observed participants selecting specific words to identify the locations on the design associated with them and selecting a subset of the markers to identify the associated impressions. This was supported by several statements:

“I wasn’t expecting the word cloud to be so helpful. Because in general when I see a word cloud I usually just assume people put words together randomly. There is no data behind it. But being able to click on each word and find individual things they cite, like the actual areas in the photo [markers] they think, enforces that. That’s really valuable.” [PL13]

“The most useful part of the interface is that I can check specific parts of the design [using visual filters]. For example, I want to see if my title in the poster works well. So I just draw a box to select this part and look at the ratings and the comments.” [PL18]
“It is the most efficient way to quickly highlight like that [draw visual filter on the design]. Oh, that part I was concerned about that. Let’s see what people thought about that part. It can help me find comments about the areas I am most interested or concerned about. I think that’s probably the best feature.” [PL14]

In addition, several participants suggested that more ways to explore the feedback would be desirable. For example, some wanted to interactively explore the feedback based on the demographics of the crowd (e.g. age and gender). This may help reveal potential patterns in the responses, aid interpretation, and lead to further insights into the design.

8.1.5 Potential Use in Practice

I shared with participants the costs and response times needed to generate feedback on their designs and discussed how this might affect the use of the system in practice. The participants did not view this as a major limitation, rather they viewed it as a fair tradeoff for not having to organize a face-to-face critique or interrupt others to receive feedback:

“No efforts from my side at all. Because I don’t need to use my social capital to do this [request feedback]. I have to call my friends at [company] and ask how my website works. They work at [company] and they are busy. But here it’s just 10 dollars” [PL17]

“I can get feedback at any time I want. For example, I did this project at 10 or 11PM at night. I’m really excited about this. I will upload it to the tool to get feedback. But it’s very inappropriate to request feedback from my professor and classmates at that time. Also, when I do freelance projects, I don’t have chance to show my designs to them. This tool allows me to work by myself.” [PL15]

Participants reported they would use a system like Voyant to receive feedback more often in the design process (Q11, μ=5.1). Participants felt the system would be particularly useful after they had begun iterating on a selected design:
“I could see myself using it in the mid to the late phase of the project. Say I have a concrete idea for the way I’m gonna approach the project. From there, after I have done a few iterations on something I was fond of. Then I may send the first or second draft down and get some feedback.” [PL14]

“When I was doing this, I was maybe about half way through the project. I think that’s a pretty good time to take data like this and look at it.” [PL16]

Some participants also wanted to use the system to generate and compare feedback on different versions of a design. For example, PL13 stated

“I could see this being incredibly useful for having multiple concepts at first. Before this, I did a quite few different sketches and concepts. I could see using the impressions on these would be really useful…. I would do three different versions of this site layout and then compare the impressions.”

This and other extensions are described next in the next chapter (Chapter 9).

**8.2 Actual Usage of Voyant in Iterative Design**

The purpose of this field study is to understand how Voyant feedback affects users’ changes to their designs and how a non-expert crowd reacts to design changes.
Figure 8.7. The designs created during the study. (a) Initial Designs: preliminary designs completed by participants in the first week. (b) Revised Designs: after receiving crowd feedback, participants spent one week revising their initial designs. Each column represents the initial / revised pair of designs for a participant.

Figure 8.8. Distribution of ratings of the depth of changes made on the Initial Designs.
8.2.1 Research Questions

Specifically, the study addresses two research questions:

RQ1: Can Voyant feedback help users improve their designs? What types and to what extent does Voyant feedback prompt changes to the design artifact?

RQ2: How effectively does the crowd notice improvements made to a design? How valid are the ratings given by the crowd?

8.2.2 Methodology

This study was conducted through collaboration with one of my thesis committee members. We applied Voyant to an entry-level course on visual design at Carnegie Mellon University. Ten participants in the course were recruited for the study, and they used Voyant in their design projects. These participants are novice designers who have not yet developed strong knowledge in visual design. This is a two-week study (see Figure 8-6). In the first week, each participant created a poster for a public event named Avant Garde (Initial Design, see Figure 8-7a). Then, participants received the feedback from Voyant. After that, participants had one more week to work on their final designs (Revised Design, see Figure 8-7b).

<table>
<thead>
<tr>
<th>participant</th>
<th>PF52</th>
<th>PF53</th>
<th>PF54</th>
<th>PF55</th>
<th>PF56</th>
<th>PF57</th>
<th>PF58</th>
<th>PF59</th>
<th>PF61</th>
<th>PF62</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Layout</td>
<td>●</td>
<td>●</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typeface</td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.2. The distribution of the four types of changes made by participants according to crowd feedback.
After participants submitted Revised Design at the end of the design iteration, they were instructed to fill out a survey, which asked them to list the changes they made based on Voyant feedback and estimate the extent of each change using a seven-point scale ranging from “Cosmetic”(1) to “Significant” (7). Participants received a $3 Starbucks gift card as compensation for completing a survey.

In addition, three experts in visual design were recruited to evaluate the quality of the collected designs. An expert rated how well a design meets each guideline and goal on a 7-likert scale. The inter-rater correlation coefficients ranged from .61 to .72, for an average correlation of .67.

### 8.2.3 Changes on Design Artifact

Most participants (N = 9) reported that they made changes based on Voyant feedback, and a majority of them (N = 6) agreed that Voyant feedback helped them make substantive changes to their designs. For example, the participant [PF59] anticipated that the first noticed element in the design is the “Avant Garde” event; however, the participant found that the first noticed element by the crowd is the building instead of the event (see Figure 8-9a). To address this problem, the participant substantially revised the design using different main images and reorganizing the visual elements. After these changes were made, the event Avant Garde became the first noticed element in the design (see Figure 8-9b).

In total, twenty-two changes were described in the survey and a majority of the changes was on the “significant” side (Avg. = 5.2, S.D. = 1.1; see Figure 8-8). The extent of the changes on a design positively correlated with the extent of quality improvement in design. \( r = 0.56, \ p < 0.05 \). The quality of a design was measured by the average of the ratings given by the experts. This result indicates that when a participant made more significant changes based on Voyant feedback, the quality of the design is more likely to be improved in the next iteration. In addition, consistent with prior work [Xu et al. 2014], the Impressions feedback was perceived as the most helpful feedback type for participants to improve their designs (Avg. = 5.5 , S.D. = 1.0).
Figure 8.9. In (a), the first noticed element by the crowd is “building”. In the revision (b), the first noticed element is the event. The feedback helped the participant [PF59] understand how to utilize an image to attract people’s attention more effectively.
The described changes on Initial Design were coded using a bottom-up approach. I found that there were four types of design changes (see Table 8-2):

- **Theme.** Change the theme and images in a design.
- **Layout.** Reorganize visual elements in a design.
- **Typeface.** Change font style and size in a design.
- **Color.** Change the color scheme in a design.

I found that significant changes were often related to the theme and layout of a design, while cosmetic changes were often related to the color and font in a design. Note that the changes listed in Table 8-2 only include the changes made based on Voyant feedback. Participants may have made additional changes which are not driven by Voyant feedback. For example, the participant [PF61] reported that he did not make any changes according to Voyant feedback. However, the participant did modify the background color in the revision (see Figure 8-7).

**Theme**

Several participants (N = 3) reported that they created new design themes to accommodate crowd responses by adopting different treatments or choices of an image in their revisions. This type of change often led to other consequential changes. As a result, these participants “pretty much redid the entire poster”. For example, one participant brightened the main image of the design, because her preliminary design yielded unintended impressions such as “hard to read” and “dark” (see Figure 8-10a):

“Received feedback about how dark the poster was... Lightened the entire poster to make it easier to read.” [PF54].

After the revision, these impression words no longer appear in the received feedback (see Figure 8-10b).

Similarly, Voyant feedback suggested another participant [PF59] to reconsider both the choice and treatment of the images:
Figure 8.10. The Impression feedback on the initial and revised designs for the participant [PF54]. (a) The impression word “hard to read” on the initial design prompted the participant to change the image treatment. The revision (b) was no longer perceived as “hard to read” by the crowd.
“But the two pictures make the article not too attractive, should have put them on another spot or other pictures maybe…”

“The pictures that grab the eye the most are just plastered on like you found them on the internet and pasted them onto a white board with glue, they aren’t integrated into the design.”

To address these problems identified in the received feedback, the participant decided to replace the images in the poster. The participant also reported that the feedback helped him understand how to utilize an image to attract people’s attention appropriately:

“The emphasis section (First Notice Feedback) helped get a sense of what was attracting the viewers’ attention, and some comments helped in getting a better sense of the overall visibility of the information or lack there of.” [PF59].

The feedback helped the participant to reconsider the treatment of the images and reorganize visual elements. The feedback on the revision indicates that the first noticed element in the design was successfully changed from the building images to the event.

Layout

A majority of participants (N = 7) revised the layout of their designs based on Voyant feedback, such as “used different grid layout for displaying the text blocks/Sections” [PF59] and “reorganized the bottom half and simplified the design” [PF53].

One participant learned that visual elements in the design were overwhelming and caused visual clutter (see Figure 8-11a). In the revision, the participant repositioned the design content to reduce the visual clutter (see Figure 8-11b):

“Making the poster much less cluttered because, again, a lot of people mentioned that… The goals section is really helpful because it lets me know if I’m doing a good job getting my point across.”’ [PF56]

Another participant reported that Voyant feedback helped him understand the effects of visual elements and the relationships among them in the design:
Figure 8.11. In (a), the design goal “visually appealing” is not well supported. After the participant [PF56] changed the layout and typeface, the average rating on the design goal in (b) was improved from 0.05 (σ = 1.5) to 1.5 (σ = 1.2). The range of the rating scale shown is -3 (low) to 3 (high).
“I knew which elements worked well with a majority of reviewers, and what elements to specifically keep”. [PF62]

As a result, the feedback helped the participant reorganize elements: “change the organization of my speakers and their [information]”.

**Typeface**

Many participants (N = 6) customized typeface styles and typeface sizes guided by Voyant feedback. Because the received feedback helped participants find out issues related to the use of typefaces in their designs, such as “font was too slanted” [PF59] and “people didn't like the color change of my fonts.” [PF57].

As a consequence, participants manipulated typefaces to address problems revealed in Voyant feedback:

> “Hanged the type of the texts to make the title standing-out and obvious... It [Voyant feedback] helps me to understand which part of my poster stands out the most, and how should I keep making such eye-capturing features in my future works.” [PF53]

> “I changed the spacing of a lot of my text. There were many comments in the impressions section about how cluttered and messy it was, so I took those into consideration when updating my design.” [PF62]

**Color**

According to the survey, participants (N = 6) attempted to better match colors to the tone they intend for their designs:

> “I changed the background color, removing the gray. People said it was overpowering and looked dreary.”[PF58] (see Figure 8-12)

> “I changed the color and gradient effect to make it more natural and realistic.” [PF55]
Figure 8.12. The Guidelines feedback on the participant’s designs [PF58]. Based on the feedback, the participant changed the layout and color in (a). The average rating on the guideline alignment in (b) was improved from -1.0 (σ = 2.3) to 1.3 (σ = 1.4).
Also, the received feedback prompted one participant to change the color scheme of the design in order to increase the contrast between the text and the background. Consequently, the readability of the content was improved:

“They find some sections hard to read against a gradient background. So I kept it consistent to one color.”[PF57]

8.2.4 Validity of Crowd Feedback

Figure 8-13 provides an overview of non-experts’ and experts' ratings of Initial and Final Designs. Non-experts gave higher ratings than experts, when evaluating Initial Designs ($F (1, 244) = 22.04, p < 0.001$). One explanation is that non-experts might have lower expectations and therefore might be less critical when evaluating a design. More interestingly, there was a statistically significant difference in the crowd’s ratings between Initial Design and Revised

![Figure 8.13. Overall quality ratings from non-experts and experts on the Initial and Revised Designs.](image)
Design (paired t-test, $p < 0.01$). This indicates that a non-expert crowd is able to react to changes made to a design.

To further understand how the crowd’s reactions are different from experts, I examined the correlation between the rating changes in the crowd and in the experts. Several steps were performed to compute the correlation. First, for a design guideline or goal, there are different numbers of individual ratings from crowd workers and experts (e.g. 20 ratings from the crowd and 3 ratings from the experts on a guideline). The average rating was used to aggregate these individual ratings. For instance, the average rating on a guideline for a design is the average of the ratings given by the 20 crowd workers. The ratings from the experts were aggregated in the same way. Second, a rating change on a guideline or goal was calculated by subtracting the average rating on an initial design from the average rating on the revised design. The distribution of the rating changes on Guidelines and Goals were show in Figure 8-14 and 8-15 respectively.

There was a moderate positive correlation between changes in Guidelines ratings from the crowd and changes in Guidelines ratings from the experts ($r = 0.45, p < 0.01$). This indicates that the aggregation of the ratings from non-experts could offer a valid assessment of improvement in a design based on standard guidelines. Moreover, I computed the correlation for each design guideline. The contrast guideline had the highest correlation value ($r = 0.59, p < 0.05$), while the alignment guideline had the lowest ($r = 0.28, p = 0.4$). One explanation is that non-experts may experience different levels of difficulty when learning different design guidelines, thus their performance varied.

However, there was no significant correlation ($r = 0.37, p = 0.07$) between changes in Goals ratings given by the crowd and changes in Goals ratings given by the experts. For example, one participant’s [PF57] design goal was to “create something that will catch the attention of people passing by” (see Figure 8-7). The experts thought the revision better satisfied the design goal, while the crowd did not agree on this assessment. One explanation is that experts and non-experts may have different ways to interpret a design goal, and thus have different perceptions of how well the design goal is achieved.
Figure 8.14. Distribution of the rating changes on Guidelines.

Figure 8.15. Distribution of the rating changes on Goals.
These results reflect the nature of design feedback, which can be conceived either as a measured judgment governed by universal principles, or personal tastes and perceptions [Venturi and Marriott 1964]. Design principles allow for a firmer basis for evaluating design, while personal perceptions are more subjective and have more variations. Voyant enables the joint use of objective and subjective data to aid in the design process. One the one hand, objective feedback such Guidelines feedback provide users opportunities to improve designs based on established design principles. One the other hand, subjective measures such as Goals feedback enable users to access different perceptions from different audiences.

In addition, rating changes from experts were more dramatic than rating changes from the crowd (see Figure 8-14 and 8-15). One explanation is that experts may have higher discriminating ability and they are more likely to react to changes in a design.

### 8.3 Content Analysis

The aim of the content analysis is to examine how the feedback generated in Voyant compares to free-form feedback. It would appear intuitive that Voyant feedback might be different from free-form feedback. However, relatively little is known about the differences.

#### 8.3.1 Research Questions

This analysis is exploratory in nature and it attempts to address three questions:

**RQ 1:** What are the genre differences between Voyant feedback and free-form feedback? Can Voyant generate feedback that goes beyond judgment?

**RQ 2:** How is the diversity of the feedback generated by Voyant and free-form format? Can Voyant offer more insights than free-form feedback by using the same amount of resource?

**RQ 3:** What are the linguistic differences between Voyant feedback and free-form feedback?
<table>
<thead>
<tr>
<th></th>
<th>Free-from Feedback</th>
<th>Voyant Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback Cost</td>
<td>2.5 dollars</td>
<td>2.5 dollars</td>
</tr>
<tr>
<td>Payment scheme</td>
<td>5 Cents per task</td>
<td>5 Cents per task</td>
</tr>
<tr>
<td>Amount of work assigned</td>
<td>Input at least 15 words</td>
<td>Input at least 15 words</td>
</tr>
<tr>
<td>Amount of time available</td>
<td>5 minutes</td>
<td>5 minutes</td>
</tr>
<tr>
<td>Number of workers</td>
<td>50 workers</td>
<td>50 workers</td>
</tr>
<tr>
<td>Number of responses</td>
<td>50 responses</td>
<td>50 responses</td>
</tr>
</tbody>
</table>

Table 8.3. Free-form feedback and Voyant feedback generated on a design.

<table>
<thead>
<tr>
<th></th>
<th>Free-from Feedback</th>
<th>Voyant Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Design is excellent looking grand. The font is very nice. I feel pleasant. very nice”</td>
<td>“The darker green text on the green. This green instantly made me think of the Statue of Liberty.”</td>
<td></td>
</tr>
<tr>
<td>“It is pretty one. The design is marvelous one. The shape of the building is very nice.”</td>
<td>“The words are not aligned properly and also the fonts are very small. The content is not clearly seen”</td>
<td></td>
</tr>
<tr>
<td>“This looks really nice ...and also color combination of black and white color so nice... and it contains total sessions and date details.”</td>
<td>“Details for each date are clearly below the date and session number, all names/details are grouped together well.”</td>
<td></td>
</tr>
<tr>
<td>“This design looks good. It highlights the name of the event, place where it is held, and the dates of the event.”</td>
<td>“It’s very colorful and clean font style. I like this page and this advertisement design also look too good.”</td>
<td></td>
</tr>
<tr>
<td>“Design is ok. It should be in light orange and the text should be in navy blue color, small text should be made readable.”</td>
<td>“Light beam is in the background. America is the world in the bounded box. Circle may be ‘o’ in the sentence in bounded box. All these elements are perfectly aligned.”</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.4. A sample of responses from free-form and structured feedback (n = 1000 in each condition).
In order to address these three questions; I performed analysis of three aspects: feedback genres, topic diversity, and psycholinguistic processes. For feedback genres, I adopted feedback categories developed by prior work and applied human coding to a sample of the feedback to identify what is discussed. For topic diversity, I used a standard Bayesian probabilistic modeling method – the latent Dirichlet allocation (LDA) to extract the topics from free-form feedback and Voyant feedback respectively. For psycholinguistic processes, I adopted the Linguistic Inquiry Word Count (LIWC) program to extract psycholinguistic features from content.

8.3.2 Data Collection

Both free-form feedback and Voyant feedback were generated on the same 20 designs (see Figure 8-7). In order to make a fair comparison with Voyant feedback from a crowd, I used the same amount of resource to generate crowd’s free-form feedback in the same conditions (see Table 8-3). For example, in free-form feedback, a worker is required to input as least 15 words. The workload of a task is similar to that of a task in Voyant’s workflows. In addition, unlike the workflows in Voyant with specific and example-based prompts, I asked the crowd a general question in free-form feedback -- What do you think about the design? Why do you think that is?

For each design, 50 responses were randomly sampled from free-form feedback and Voyant feedback respectively (see Table 8-4). The cost of these responses is the same as that of the crowd’s free-form feedback. As a result, the data set includes 1,000 responses in each condition.

8.3.3 Feedback Genres

As Voyant is able to generate five types of feedback, I first examined whether free-form feedback can generate these feedback types. Based on the definition of the feedback types described in Chapter 5, I coded a random sample of 100 responses in free-form feedback. I found that very few responses discussed First Notice, Guidelines, and Goals. Specifically, only 4 responses described first noticed elements in a design; 6 responses commented on
design guidelines; 8 responses discussed the intention of a design. All these feedback types are desired from non-experts. However, free-form feedback did not provide these desired feedback types. Having specific prompts in feedback generation is therefore important.

In order to comprehensively assess the category differences in free-form feedback and Voyant feedback, I coded the collected feedback using categories derived from a typological analysis [Dannels and Martin 2008], in which nine feedback categories were developed from the design feedback given by both experts and novices. 100 responses were randomly sampled from both free-form and Voyant feedback. The unit of analysis was any sentence or phrase which represented a coherent statement about a design. The responses generated 260 units for coding and each feedback unit was coded according to feedback type. The units were independently coded by two coders (kappa = 0.80) and disagreements were solved through discussion. Table 8-5 shows the frequency of occurrence of each category in free-form and Voyant feedback. Several categories such as comparison, identity invoking and process oriented feedback did not occur in the collected feedback. One explanation is that these categories were derived from experts’ feedback [Dannels and Martin 2008] and non-experts were not able to provide insights into users’ design approaches and contemporary trends.

Judgment is the most significant feedback category ($\chi^2 = 15.7, p < 0.001$; see Table 8-5). The judgment category occurred more often within free-form feedback (46%) and much less often in Voyant feedback (32%). Workers in the free-form condition offered many simplistic judgments, especially related to positive reactions such as “I like it”. This result corroborates

<table>
<thead>
<tr>
<th></th>
<th>Free-form Feedback</th>
<th>Voyant Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Judgment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>86%</td>
<td>78%</td>
</tr>
<tr>
<td>Negative</td>
<td>14%</td>
<td>22%</td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual</td>
<td>69%</td>
<td>40%</td>
</tr>
<tr>
<td>Relation</td>
<td>31%</td>
<td>60%</td>
</tr>
<tr>
<td><strong>Suggestion</strong></td>
<td>12%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Table 8.5. Frequencies of the categories of critique discourse found in each form of feedback.
previous findings that a majority of online feedback does not go beyond simplistic judgments [Willett, Heer and Agrawala 2012; Xu and Bailey 2012].

Feedback coded as interpretation occurred when crowd workers tried to make sense of a design and explain their perceptions of it. Interpretation was the most common category in Voyant feedback (64%). Voyant feedback had higher frequency of interpretation than free-form feedback ($\chi^2 = 5.8, p < 0.05$). In Voyant, after workers describe their perceptions, they perform a related task to explain these perceptions. This separation could help workers pay more attention to the interpretation process.

During the interpretation process, crowd workers in Voyant more often associated their perceptions with elements in a design than workers did in free-form feedback (see Table 8-5). Workers in free-form feedback attempted to associate their perceptions with individual elements:

“The tall building gives you the idea of a skyscraper”.

In contrast, a worker in Voyant was more likely to discuss relations between several elements in a design:

“Having the text on the building helps establish a spatial relationship in an attractive manner.”

One explanation is that workers in Voyant feedback are exposed to multiple specific design elements (e.g. design elements offered by the system to help a work explain the rationale). These specific examples could facilitate the interpretation process and help workers consider the relation among elements in a design.

The suggestion category was also significant ($\chi^2 = 7.4, p < 0.01$). Free-form feedback offered more suggestions for improvement than Voyant feedback (see Table 8-5). Although crowd workers were not rewarded for offering suggestions in either condition, workers in free-form feedback provided more suggestions. One explanation is that the open nature of free-form gives people license to suggest improvements. I also observed that suggestions offered in the feedback mainly recommend users to make trivial changes to their designs. For example, a crowd worker wrote in the free-form condition:
“Background should be in light yellow color. Text and log image should be in dark blue color.”

The quality and effect of these suggestions needs further investigation.

### 8.3.4 Topic Diversity

I used LDA to extract the topics from free-form and structured feedback [2]. LDA is a standard topic modeling method that is often used to discover hidden topics in documents and the words associated with each topic [2]. The method excels at analyzing large amounts of unlabeled documents by clustering words that frequently co-occur.

Several steps were performed to process the data. First, clean the data. I converted the text to lowercase and removed the punctuation characters (');:.“”. Second, create “Bag of words”. I adopted a unigram “bag of words” model, a common approach in computational linguistics [Manning and Schütze 1999]. That is, I used single words (unigrams) to represent a response in the feedback data.

Based on the preprocessed text, I built LDA topic models from free-from feedback and Voyant feedback respectively, treating each response as a document. In order to make a fair comparison, the number of topics was set as the same in the models. The number of topics was tested between five and twenty, and the differences between from free-from feedback and Voyant feedback were observed consistently. Thus, I only report the results obtained with the number of topics set to fifteen (see Table 8-6).

The resulting topics in free-form feedback had much (i.e. many topics shared the same word “design” and “like”). It appears that Voyant feedback offered a wider range of topics (see Table 8-6). I further computed the cosine similarity between any of these two topics in free-form feedback and Voyant feedback respectively. The similarity values among topics were visualized by a heatmap (see Figure 8-16). I observed that the topics in Voyant feedback were more diverse than the topics in free-form feedback (Student's t-test, \( p < 0.01 \)). This observation probably can be explained by the use of various prompts in tasks of generating Voyant feedback.
Moreover, Voyant feedback provided more detailed information than free-form feedback. In free-form feedback, crowd workers often referred to “design” and subjective judgments such as “like”, “nice”, and “great” (see Table 8-6). For example, one response was:

### Table 8.6. A standard topic modeling method LDA was applied to detect the latent topics in the feedback. There were 15 topics in the feedback and the top 5 topic words were listed for each topic.

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Free-form Feedback</th>
<th>Voyant Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>like, good, design, nice, font</td>
<td>look, goal, impression, work, form</td>
</tr>
<tr>
<td>Topic 3</td>
<td>design, color, text, font, think</td>
<td>first, word, easily, attract, letter</td>
</tr>
<tr>
<td>Topic 4</td>
<td>design, background, color, font, nice</td>
<td>text, section, headline, contrast, attraction</td>
</tr>
<tr>
<td>Topic 5</td>
<td>design, like, good, contain, avant</td>
<td>design, look, attractive, shape, overall</td>
</tr>
<tr>
<td>Topic 6</td>
<td>color, background, easy, sir, good</td>
<td>font, style, title, size, color</td>
</tr>
<tr>
<td>Topic 7</td>
<td>design, nice, attractive, good, building</td>
<td>like, image, nice, clear, attractive</td>
</tr>
<tr>
<td>Topic 8</td>
<td>building, nice, great, design, font</td>
<td>good, read, structure, highlighted, match</td>
</tr>
<tr>
<td>Topic 9</td>
<td>design, looks, font, nice, white</td>
<td>information, font, clearly, easy, written</td>
</tr>
<tr>
<td>Topic 10</td>
<td>design, font, text, building, like</td>
<td>color, look, great, beautiful, combination</td>
</tr>
<tr>
<td>Topic 11</td>
<td>design, like, color, background, image</td>
<td>building, page, attractive, tower, help</td>
</tr>
<tr>
<td>Topic 12</td>
<td>good, design, like, building, nice</td>
<td>poster, line, date, event, clearly</td>
</tr>
<tr>
<td>Topic 13</td>
<td>good, design, like, nice, color</td>
<td>background, color, white, black, red</td>
</tr>
<tr>
<td>Topic 14</td>
<td>detail, grade, america, like, date</td>
<td>different, people, make, attractive detail</td>
</tr>
<tr>
<td>Topic 15</td>
<td>color, like, read, good, nice, poster</td>
<td>architecture, massive, logo, new, catch</td>
</tr>
</tbody>
</table>
“The design looks very nice. So I like it very much. Awesome design to see.”

In Voyant feedback, workers focused on more specific issues such as “elements”, “structure”, “layout”, and “style”. The results suggest that tasks in Voyant help workers attend to specific design elements rather than just provide holistic judgments. For example, one worker wrote:

“The darker green text on the green. This green instantly made me think of the Statue of Liberty.”

The worker not only paid attention to specific elements in a design, but also used these elements to explain his perceptions of the design.

8.3.5 Linguistic Styles

LIWC was adopted to measure linguistic styles for each response in both Voyant and free-form feedback. LIWC, a dictionary developed in the psycholinguistic field, has been used extensively in psychological and HCI research to quantify the grammatical, linguistic, and psychological features of text documents. LIWC includes 70 pre-defined categories, and dictionaries of words related to each category. I hypothesized that these language-use categories might be related to the differences between free-form feedback and Voyant feedback.

First, I used LIWC to analyze language use in each response both in free-form feedback and Voyant feedback. Specifically, for each response, I applied LIWC to calculate the degree to which people use a category of words across texts. In other words, each response in the feedback data is measured by percentages of the total words belonging to specific categories.

Second, a series of t-test comparisons were carried out between free-form feedback and Voyant feedback in terms of the LIWC generated language-use categories. As described before, my hypothesis is that free-form and Voyant feedback are equal in the LIWC features. The null hypothesis is considered to be rejected at \( p < 0.05 \). All the categories in LIWC are considered. Due to these 70 simultaneous tests, I allow for a Bonferroni correction, letting \( \alpha = 0.05/70 = 0.00071 \). Table 8-7 shows the average and standard deviation of proportions of words in LIWC categories along with \( p \)-values associated the corresponding t-test.
Figure 8.16. A heatmap visualizes the similarity values between any two topics. The lighter color means the similarity value is higher. The topics in Voyant feedback are more diverse than the topics in free-form feedback.
Linguistic Processes

Voyant feedback had more words per sentence, more long words (6-letters or greater), more impersonal pronouns such as "those" and “it”, more prepositions, more articles, more total function words, fewer auxiliary and common verbs, and fewer words related to future tense and present tense (see Table 8-7). A higher ratio of words per sentence and long words is often used as an indicator for better domain-specific working knowledge [Kamhi and Catts 2012; Kim et al. 2011]. One interpretation is that workflows in Voyant can facilitate a non-expert crowd’s ability to interpret a design and communicate their perceptions of a design, whereas free-form comments may inhibit non-experts’ potential to develop thoughtful responses in the discussion of creative work.

In addition, Voyant feedback has a lower rate of auxiliary verbs and a higher rate of articles, which are strongly related to making good grades in education. Previous work also shows that people who use articles at a high rate tend to be categorical thinkers, who are more concrete in their thinking [Kamhi and Catts 2012]. One possible explanation is that workflows in Voyant can enhance categorical thinking and thus help crowd workers focus on specific objects and categories.

Perceptual Processes

Differences between Voyant feedback and free-form feedback were found in the use of words related to feeling, a subcategory of perceptual processes (see Table 8-7). Crowd workers in Voyant feedback are more likely to express perceptual process words such as feel, relative to when they are in free-form feedback. The higher use of feeling words in Voyant feedback could be caused by the tasks in Voyant’s workflows, which explicitly and repeatedly instruct workers to explain their perceptions of a design. These instructions could encourage crowd workers to provide personal experiences and information in order to explain their perceptions.

Affective Processes

Significant differences were found for affective processes (see Table 8-7). Crowd workers in free-form feedback used more emotional words than did crowd workers in Voyant feedback.
In many domains, emotion regulation is reported to be a part of expertise [Chaffin and Imreh 2002; De Groot and Groot 1978]. As described previously, workers in both free-form and Voyant feedback did not have domain-specific knowledge. One explanation is that when non-experts are asked to provide spontaneous judgments of a design, they attempt to reply on affective processes and analyze a design in an emotional manner. In contrast, specific examples and prompts in tasks of Voyant feedback may inspire crowd works to analyze a design rationally instead of solely replying on their affective processes.

**Cognitive Processes**

I found that words related to certainty (e.g. always, never) were more dominant in Voyant feedback (see Table 8-7). Increasing the use of certainty words is an indicator of improved critical thinking [Carroll 2007]. It is possible that feedback tasks in Voyant can facilitate the development of critical thinking thorough specific examples (e.g. listing top three design elements identified by previous workers).

The analysis revealed that free-from feedback had more discrepancy words (e.g., should, would, could) than Voyant feedback (see Table 8-7). This result is consistent with the prior finding that crowd workers in free-form feedback are more likely to offer suggestions and solutions than workers in Voyant feedback. Voyant mainly focus on helping users to find problems in users’ designs. Finding problems is important. Because if users want to improve their designs, they must first understand what problems they want to fix. However, in the future, it would be interesting to explore how to utilize fee-form feedback to brainstorm solutions to fix the identified problems.

**Spoken Categories**

The analysis showed that the spoken language usages were significantly different between Voyant feedback and free-form feedback (see Table 8-7). Free-form feedback was characterized by more spoken categories including more filter and assent words. This observation indicates that Voyant feedback is more formal and deliberate. One explanation is that feedback tasks in Voyant purposely reduce spontaneous judgments and lead crowd workers to focus more on the design than on crowd workers themselves.
<table>
<thead>
<tr>
<th>Linguistic Processes</th>
<th>Free-form Feedback</th>
<th>Voyant Feedback</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Auxiliary Verbs</strong></td>
<td>10.26 (6.68)</td>
<td>7.80 (5.90)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Words/Sentence</strong></td>
<td>15.76 (7.52)</td>
<td>18.50 (7.14)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Common Verbs</strong></td>
<td>14.35 (7.25)</td>
<td>11.92 (6.90)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Words&gt;6 letters</strong></td>
<td>16.56 (8.85)</td>
<td>19.01 (10.12)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Future Tense</strong></td>
<td>1.18 (2.75)</td>
<td>0.63 (2.02)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Prepositions</strong></td>
<td>8.92 (6.01)</td>
<td>10.13 (6.06)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Total Function Words</strong></td>
<td>51.63 (9.81)</td>
<td>53.67 (10.55)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Impersonal Pronouns</strong></td>
<td>5.63 (5.12)</td>
<td>6.73 (5.90)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Present Tense</strong></td>
<td>10.16 (6.08)</td>
<td>9.04 (5.89)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Articles</strong></td>
<td>10.25 (6.52)</td>
<td>11.55 (6.87)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Total Pronouns</strong></td>
<td>7.78 (6.72)</td>
<td>8.80 (7.18)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Perceptual Processes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Feel</strong></td>
<td>0.35 (1.31)</td>
<td>0.59 (2.00)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Affective Processes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Affect Process</strong></td>
<td>9.34 (7.32)</td>
<td>6.75 (6.54)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Positive Emotion</strong></td>
<td>8.86 (7.24)</td>
<td>6.37 (6.53)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Cognitive Processes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Discrepancy</strong></td>
<td>1.32 (2.93)</td>
<td>0.62 (2.04)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Certainty</strong></td>
<td>1.16 (2.39)</td>
<td>1.51 (2.82)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Social processes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Social</strong></td>
<td>1.87 (3.79)</td>
<td>3.35 (4.45)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Relativity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Space</strong></td>
<td>5.24 (5.66)</td>
<td>6.26 (6.24)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>3.41 (4.51)</td>
<td>2.80 (4.13)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Spoken Categories</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fillers</strong></td>
<td>0.72 (1.87)</td>
<td>0.29 (1.23)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Assent</strong></td>
<td>0.57 (1.69)</td>
<td>0.26 (1.22)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Punctuation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Commas</strong></td>
<td>1.60 (3.69)</td>
<td>3.50 (10.75)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Quotation Mark</strong></td>
<td>2.68 (3.79)</td>
<td>3.41 (4.11)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Colon</strong></td>
<td>0.002 (0.06)</td>
<td>0.067 (0.61)</td>
<td>&lt;.05</td>
</tr>
<tr>
<td><strong>Exclamation Mark</strong></td>
<td>0.081 (0.67)</td>
<td>0.012 (0.22)</td>
<td>&lt;.05</td>
</tr>
</tbody>
</table>

Table 8.7. The average and standard deviation of proportions of words in LIWC categories along with \( p \)-values associated the corresponding t-test (\( n = 1000 \)).
8.4 Summary

Three studies were conducted to evaluate Voyant from three perceptive: users’ perceptions, influence on design solutions, and feedback content.

The lab study examined users’ perceptions of the benefits of Voyant feedback, the usability of the visual feedback representations, and interactions in the interface. Participants reported that Voyant feedback was helpful, especially the feedback about impressions and interpretations of their goals. The feedback helped participants identify design problems, gain insight into their designs, and consider their designs from diverse perspectives. Also, the interactive coupling provided by Voyant was helpful for analyzing the association between the crowd’s perception of a design and the elements within it. Moreover, participants reported that Voyant feedback was a useful complement to prior feedback received on their designs.

The second study investigated how Voyant feedback affects users’ designs in practice. Most participants reported that Voyant feedback helped them improve their designs and a majority of the changes was on the significant side. Significant changes were often related to the theme, layout of a design, while cosmetic changes were often related to the color and font in a design. When a participant made more significant changes according to Voyant feedback, the quality of the revision is more likely to be improved. In addition, the study showed that a non-expert crowd with appropriate instructions is able to offer a valid assessment of the value of a design based on standard design guidelines.

As compared with free-form feedback, Voyant feedback produced more interpretative, diverse and critical discourse. I performed content analysis of three aspects: feedback genres, topic diversity and psycholinguistic processes. First, interpretation was the most common category in Voyant feedback, while judgment is the most common category in free-form feedback. Surprisingly, free-form feedback offered more suggestions than Voyant feedback. Second, the topics in Voyant feedback were more diverse and detailed than the topics in free-form feedback. Finally, linguistic style differences indicate that specific examples and prompts in Voyant help crowd workers develop critical thinking. Structured workflows in Voyant purposely reduce spontaneous judgments and lead crowd workers to adopt various analyzing strategies instead of solely replying on their affective processes.
Chapter 9
Discussion and Future Work

In previous chapters, I presented how to utilize an online crowd as a simulated audience to generate perception-oriented feedback on visual designs. A new system, Voyant, was implemented to enable users access to a non-expert crowd in order to assess how the crowd perceives different aspects of the design.

In this chapter, I discuss some of the broader issues related to the system and propose several natural extensions, which could generate more helpful, affordable, and timely feedback using online crowds. I also discuss directions that extend crowdsourcing feedback on visual designs, for applying crowdsourcing techniques in different creative domains, and for supporting creative activities from the individual to the collective.

9.1 Finding Problems
My system helps users identify problems with their designs, but it is up to the users to determine how to best modify the designs to address those problems. The fact that the feedback generated by the system can surface problems with a design is itself a significant outcome, e.g., as iterating on a design requires knowing what problems to address [Dow, Heddleston and Klemmer 2009]. The user can also inspect the overviews, visual markers, and explanations to not only identify problems but also understand why they occurred. The feasibility of using crowds to brainstorm solutions to the problems identified could be explored in future work. It would also be interesting to observe how Voyant is utilized throughout an entire design project. Such a study could reveal how the feedback benefits different design phases and types of designs, and users with different levels of design expertise.
9.2 Improving Feedback Quality

The *Guidelines* feedback received ratings lower than the other types. One cause could be that the worker tasks for *Guidelines* were more complex than the others. For instance, when rating a design relative to a guideline, a worker had to comprehend the guideline, review the examples, and apply it to the design. Outcome quality may therefore have been more variable on these tasks. To improve quality, additional scaffolding for workers such as apprenticeship or showing worked examples for the tasks could be explored. For the other feedback, only minimal quality control was performed and participants in the evaluation did notice that some comments were irrelevant, which cast doubt on the corresponding input. Voyant and similar systems should therefore consider adopting additional quality controls such as use of self-assessment [Dow et al. 2012], find-fix-verify [Bernstein et al. 2010], validation questions [Kittur et al. 2008], or task designs [Kittur et al. 2013] to improve feedback.

Participants in our studies appreciated exploring the feedback through the visual markers placed on their designs. Each marker was placed by a crowd worker to associate an element of the design with her perception during a task. Placing a marker was typically straightforward, but became problematic when the element could not be easily localized, e.g., background color, white space, or relations between multiple elements. System designers could improve this aspect by offering the worker different types of markers, e.g., different shapes, free-form annotation, and selectable areas or groups defined by image segmentation techniques.

9.3 Measuring Feedback Reliability

Another area that deserves attention is the reliability of the feedback. Voyant like many crowd-based systems allows users to reach a crowd through an open-call format. If a user requests feedback at different times, the crowd will be composed of different workers and thus the received feedback might be different. For example, Figure 9-1 shows the feedback on the same design generated at two different times. Large impression words in the word clouds remain the same; however, most of the smaller impressions are different. The current system does not provide an explicit way for users to assess the reliability of the feedback.
Figure 9.1. Voyant generated the Impressions feedback on the same design at different times. Frequent impressions such as “colorful” and “juice” in the word clouds remain the same, but less frequent impressions are different in (a) and (b).
The concept of feedback reliability is complex. The main challenge is that the desirability of feedback reliability depends on the context of a specific design. For example, consistent feedback may be welcome for designs that need to communicate clear and consistent messages, whereas uncertainty or controversy in feedback could be desired for designs that are intended to evoke affective responses. As a first step, crowd feedback systems should provide straightforward statistical measures such as variances of the distributions of worker responses to help users assess the feedback reliability. Beyond numeric measures, future systems could help users manage the reliability of crowd responses by collecting small batches of worker responses. This approach would allow users to receive a small amount of feedback and then decide whether they want to receive additional responses either as a means for soliciting new viewpoints or as a confirmation of the feedback already received. Moreover, experience sampling methodology could be adopted to collect ground truth data for understanding and modeling the perceived level of reliability from users’ perspectives [Csikszentmihalyi and Larson 1987].

9.4 Optimizing Time and Cost

Our studies revealed that users may have different expectations about the feedback generation speed. For example, some users want to get the feedback in two hours, while others are willing to get the feedback in two days. In the future, a system could allow users to set a deadline for the feedback and the system could dynamically adjust the task price and the workflow based on the deadline.

Future work could also explore how to optimize the workflows in terms of the workers, costs, and time required. For example, after a worker learns a design guideline and completes the corresponding feedback task, future systems can recommend the user to continue to evaluate other designs based on the same guideline. Moreover, future systems can suggest relevant feedback tasks based on a worker’s performance and interests. Appropriate recommendations may encourage workers to keep working on feedback tasks [Cosley, Frankowski, Terveen and Riedl 2007]. As a result, the task completion rate can be increased, and the response time and the cost of the feedback can be reduced.
9.5 Supporting Comparison of Feedback on Multiple Designs

The current system generates feedback on one design at a time. However, if desired, a user could generate feedback on multiple variations of a design or competing concepts. The feedback could be used to conduct deeper comparison between the designs and complement simple AB tests (e.g. as offered in [Bernstein et al. 2011; Dow, Gerber and Wong 2013]). Users may want to compare the feedback of different designs for various reasons. Ultimately, the reasons they discussed were all tied to a desire to take action to improve their present work.

A variety of visual analytics systems have been developed to support a comparison of a history of creative activities. As an early project in this space, Netscan [Brush et al. 2005; Smith 2002] offered visual dashboards for Usenet users to track metrics and contents in newsgroups. Netscan provided a comparison of various Usenet groups so that people could decide which groups they might want to join based on their topic or activeness. More recently, history Flow [Viégas et al. 2004] and WikiChanges [Nunes et al. 2008] visualized a history of content changes in Wikipedia articles to show the evolution of article content over time. Such visualizations have helped to enhance our understanding of the dynamics of creative activities.

A related extension is to visualize the feedback history of the evolution of a design. For example, the feedback could be shown and made accessible through an interactive, visual timeline [Dunne et al. 2012]. The system also could allow users to bookmark informative views of the feedback as a way to help guide and explain later design choices [Heer, Viégas and Wattenberg 2007].

9.6 Reflections on Use Cases and Users

An underlying assumption of my work is that crowd feedback is intended to aid a class of users who want to receive audience feedback in order to improve their designs. However, according to observations reported in Chapter 8, there could be a number of unintended uses of crowd feedback. In design education, the crowd feedback on a design could be used by students to argue for a higher grade or to counter the recommendations provided by the
instructor or peers. Also, crowd feedback could contribute to facilitating communication between designers and clients. For example, crowd feedback could be used as an argument and rationale to help designers persuade clients about specific design choices. Clients could also leverage crowd feedback to better specify their requirements.

Crowd feedback systems enable users to receive feedback from an external audience. The feedback is more valuable for designs aiming to deliver clear messages that the audience can understand than for designs intended to be more artistic and less functional. Also, crowd feedback systems could be more beneficial for novice designers and non-designers who cannot receive desired feedback from their professional or social networks.

Another consequence is that crowd feedback systems do not necessarily enhance creativity in the design process. Designers should be cautious that the feedback-driven approach overly controlling design decisions may inhibit their own creativity.

9.7 Other Feedback Types and Crowds

Five types of feedback are generated in Voyant, but other types may also be beneficial. For example, the system could enable dialog between the user and the crowd, e.g., by allowing questions to be posed for each other about the design. Other feedback such as the analysis of a design relative to precedents or contemporary trends [Dannels and Martin 2008] would also be useful, but cannot yet be generated. To the degree that a crowdsourcing or other computational approach could generate these types of feedback is an open question.

The current system recruits a crowd of workers from an online labor market, but other crowds such as the employees of a large organization, members of an online community, or students in a large design course could also be used. In the latter case, the perception-oriented feedback generated by the system could complement other forms of peer critique already integrated into the course [Kulkarni and Klemmer 2012]. A related future study would be to compare the feedback generated by crowds with different levels of design knowledge.
9.8 Other Studies of Free-form Feedback and Structured Feedback

The analysis of feedback content in the structured and free-form conditions showed that crowd workers suggest more solutions when prompted to enter free-form feedback. Note that workers were not asked to propose solutions in either condition. One explanation is that an open response format does not constrain the range of feedback provided by the workers. In contrast, using structured workflows appears to cause workers to respond only as instructed in the task and thus did not propose solutions. Alternative workflows could be proposed to provide suggestions for improvement. For example, when workers give low ratings on a design’s communicative goal, the system can prompt these workers or recruit additional workers to brainstorm suggestions for improving the design with respect to that goal. Future empirical studies need to be conducted to examine the effectiveness of these workflows.

The differences between free-form feedback and structured feedback should be further evaluated. My current studies show that structured workflows help non-experts provide meaningful feedback. It is interesting to examine how experts perform in both free-form and structured feedback conditions. In addition, in the studies, I only investigated the content differences between Voyant feedback and free-form feedback. For example, I found that Voyant feedback is more diverse and detailed than free-form feedback. However, I did not examine how these content differences can affect users’ design outcome. A 2×2 factorial design study can be conducted to address these questions. The two factors would be: the feedback format (structured feedback or free-form feedback) and the expertise level of a crowd (non-experts or experts). In this study, users will be randomly assigned into the four conditions. In each condition, users will receive the corresponding feedback on their designs in the design process. The study will allow understanding of how different feedback formats and expertise levels of a crowd can help users iterate toward a design solution.

9.9 Crowdsourcing Feedback in Other Creative Domains

Voyant targets the visual design domain. This domain was selected because it is challenging and more people are performing visual design work due to the increased demand for visual communication skills and the wide availability of graphic design software. But, the concept of
receiving structured feedback from non-experts may be applicable to other domains. The following are several examples of possible application areas: crowdfunding, architecture, and medical education.

Crowdfunding offers a new paradigm for entrepreneurs to initiate, expand, or advertise their business ideas [Gerber and Hui 2014; Muller et al. 2013; Xu et al. 2014]. Many campaigns have succeeded in reaching their funding goals, however, more than half of the campaigns have failed [KickstarterStats]. As pointed out by Gerber and Hui [Gerber and Hui 2014; Hui and Gerber 2012], publishing crowdfunding campaigns consumes a great amount of users’ social capital. Users need to spend a lot of time interacting with crowds (funders) and advertise their campaigns. People might not be likely to fund the same campaign more than once. Users need to ensure that the campaigns are likely to reach the funding goal before they publish the campaigns on a platform. Crowdfunding sites such as Kickstarter and Indiegogo provide guidelines and examples of past campaigns to help creators build successful campaigns; however, there are no existing methods that can help users assess and improve the effectiveness of their campaigns prior to launch. A system like Voyant may therefore be helpful for users to receive quality feedback before they launch their campaigns.

In building architecture, a system like Voyant could enable architects to receive impressions of the form of a building or interpretation of its design goals. Leveraging virtual or augmented reality, we could enable crowd workers to perform the 3D walk through and annotate their experience or interpretations of the space. Finally, the concept of receiving crowd feedback could be applied to medical education. For instance, a crowd can be used as simulated patients to help aspiring physicians learn how to better communicate with patients.

9.10 Hybrid Combination of Crowd and Machine

Generating feedback is complex. Human intelligence and machine intelligence both have their own advantages. On the one hand, it is undeniable that unstructured information, rare patterns, and information that requires high levels of semantic interpretation are still difficult to be processed by machines [Dawes et al. 1989]. In reality, we also do not have the correct formulas for every desired feedback type [Kleinmuntz 1990]. On the other hand, mathematical models can ensure that the variables contribute to the result based on actual predictive power.
Thus, models often outperform humans when the judgment requires integration of multiple variables [Dawes, Faust and Meehl. 1989].

The feedback in Voyant is purely made by humans. Can machine intelligence be integrated into the feedback generation process? Can certain types of desired feedback be generated by machine learning techniques? For example, we could utilize computer vision techniques to automatically detect visual elements in a design. These identified elements can help crowd workers explain their perceptions of a design. Also, future systems can assess worker input in real-time and create customized examples that aid the feedback generation process.

The feedback generation process could also be iteratively optimized. First, one can collect human feedback on individual attributes and the predefined tasks can be applied to collect human feedback from crowd workers. Then, based on the analysis of the collected feedback, new tasks can be dynamically created to prompt workers to provide feedback on additional attributes and collect insights that are not obtained in the existing feedback data. Finally, in order to ensure the readability of the feedback, mathematical models can be proposed to synthesize the content and improve the consistency of writing style of the feedback.

9.11 Example Search based on Crowd Feedback

The feedback generated by Voyant can be treated as a labeled data set. I would like to utilize advanced techniques in natural language processing and information retrieval to develop a crowd-based example search system, which could enhance how people search design examples in their creative practice. The Web is widely used to search and find examples in a user’s creative practice. While today’s systems generally allow queries on HTML metadata and image content, these systems cannot support queries based on how an audience interprets a design.

The structured feedback generated by a crowd feedback system like Voyant can be treated as a labeled data set based on an audience’s perceptions. The data set includes impression word clouds that describe the design and visual markers that point to elements that go with those words. Imagine that we have thousands of designs stored in a repository with all of the feedback provided. Suppose a new user wants to search for examples that convey the
impression of *happy* from a target audience. A future system would allow the user to search the repository for images based on an audience’s impressions. The system could also show how the *happy* impression was formed and evolved through iterations.

This capability could be especially useful in design education. For instance, design examples retrieved based on the semantics of a design concept can help students understand how to better communicate intended messages through visual elements. Moreover, future system could help instructors provide their students with learning examples.

### 9.12 Managing Alternatives on a Large Scale

The current implementation of Voyant focuses on individual designs. It works well when the number of design alternatives is relatively small. The comprehensive feedback offered by Voyant can help users refine their designs iteratively.

However, when users have a large number of alternatives generated by a crowd, they need to first filter promising alternatives and then further refine the selected ones. For example, a crowdsourcing design website like Crowdspring or 99designs allows a user run a design competition and then designers from all over the world submit their designs. A user only needs to pay several hundred dollars and can receive hundreds of visual designs (e.g. logos and posters) within a week [Crowdspring]. Voyant is not effective in this case, because the user is not the creator of the designs and s/he needs the feedback to filter rather than improve these designs. Also, it is impractical for the user to review all the feedback on hundreds of design alternatives.

Future work could develop systems for helping users select promising ideas from a large number of ideas generated by a crowd. For example, a system could recruit crowd workers to score a design based on certain criteria. Scalable scoring methods can be applied to aggregate the individual scores for idea selection [Xu and Bailey 2012]. These future systems can assist users in quickly narrowing a large space of ideas to a smaller subset for closer scrutiny.
9.13 Social and Cultural Factors in Crowd Feedback Systems

A crowd feedback system is not only an economic system, but also can be a sociological one. The question is how many social and cultural issues should be brought into the design of crowd feedback systems. Here I list four issues to consider for future work.

**Social Capital.** In traditional face-to-face settings, engaging in critique activities helps users build and maintain social capital. Continuous communication among users also helps them become mutually familiar with each other's work, and the knowledge of previous iterations on designs helps them provide contextualized feedback. However, this is lost entirely in crowd feedback systems. Current systems provide a stream of anonymous crowd workers working on a users’ design, and these workers may never work for the same user again. Alternative crowdsourcing mechanisms could be proposed to allow users to recruit the same workers to provide feedback on subsequent iterations of the design. For example, one approach would be to allow the user to mark feedback responses that were especially helpful and the system would recruit the corresponding workers for a higher wage or bonus. Future work is needed to assess the desirability and use of this type of feature in a crowd feedback system.

**Reputation System.** Voyant currently recruits crowd workers from Amazon Mechanical Turk, which has a built-in performance-based reputation system for workers. For example, users can require workers with a minimum level of worker's HITs approval rate. Yet, such a reputation is not likely to be sufficient to insure that users can receive quality feedback on visual designs. The current system also allows users to filter workers based on the workers’ demographic information; however, many users want to learn more about workers’ background information (e.g. workers’ activeness and educational background) because the information could help them better weigh the feedback. Having design-relevant reputation information about workers would allow users to select desired workers and help users understand the feedback. Additional reputation information could be obtained by allowing users to rate and comment on workers’ responses and aggregating these ratings and comments across all users of a crowd feedback system.

**Motivation.** Current crowd-based systems are dominated by extrinsic motivation [Mason and Watts 2010]. The high financial incentive could make workers accountable for their actions,
but it also could undermine workers’ intrinsic motivation. Additional motivation mechanisms can be integrated into crowd feedback systems. For instance, the intrinsic motivation of workers can be fun of working, enjoyment of helping, and the feeling of making a unique contribution. Future studies could be conducted to understand the effects of different types of motivation on various aspects of performance. For example, it would be interesting to examine how feedback from workers in an unpaid condition, where workers perform the work ostensibly for enjoyment, compares to feedback from workers incentivized with micro-payments or other benefits.

**Domain Expertise.** Crowd workers may have different levels of expertise on various subjects. For example, workers may have different expertise in critiquing ideas and creating ideas. In order to help users screen for expertise, future work could empirically evaluate different expertise measures (e.g. Bayesian Truth Serum [Prelec 2004], Conventional Wisdom Score [Bollen and Lennox 1991], and test of design knowledge [Luther et al. 2014]) and apply effective ones to a crowd feedback system. Moreover, a visual analytics interface can integrate this information and allow users to interactively filter and explore the feedback based on workers’ expertise level.
Chapter 10

Conclusion

Feedback is a critical aspect of design. However, existing methods for receiving feedback mainly depend on a user’s social capital and peer networks. The access to desired feedback is limited. An emerging opportunity for receiving feedback on designs is harnessing a non-expert crowd. My dissertation has made the following contributions to advance the science of crowd feedback systems:

One contribution of my dissertation is the results of a formative study revealing several types of feedback on visual designs that are desirable from non-experts. This contribution is significant because my results point to feedback (e.g. interpretation of the user’s goals) that is not directly provided in prior research or existing systems. The results can also provide justification for creating or using crowd feedback systems such as Voyant.

The second contribution is the design and implementation of Voyant – a new system to enable access to non-expert crowds in order to assess how a simulated audience perceives different aspects of the design. Voyant exemplifies how crowdsourcing can be harnessed to enable new forms of computational tools that support design activity. The key features of the system include generating structured feedback on visual designs from non-experts and providing coordinated, interactive views for exploring the feedback. End users can leverage Voyant to iterative toward a more effective solution. Voyant can also be used by researchers as an experimental platform to study the effects of crowd feedback in a longitudinal context.

The third contribution is the evaluation results of the feedback generated by the system for authentic designs and the effectiveness of the interactions. Specifically, my evaluations revealed that crowd feedback was helpful for users to improve their designs in practice, especially the feedback about impressions and interpretations of users’ goals. The feedback helped users identify design problems, gain insight into their designs, and consider their designs from diverse perspectives. Users also appreciated the coordinated views and the
interactions for exploring and understanding the feedback generated. As compared with free-form feedback, Voyant feedback produced more interpretative, diverse and critical discourse.

In conclusion, my dissertation has explored the use of non-expert crowds to generate helpful, affordable, and timely feedback for visual design. The promising findings of this work will hopefully motivate the research community to continue to make progress on this topic of crowd feedback. This could have a large impact on design practice. Users will have more opportunities to make design decisions based on audience input rather than solely on intuitions. It is pushing toward a future of data-driven creativity. It will be interesting to see which aspects will be embraced and which aspects will be resisted by the design community.
Appendix A: Consent Form

You are invited to participate in a research study that seeks to understand the difference and similarity among critiques received from experts, critiques received from online communities, and critiques received from the crowdsourcing-based approach. The knowledge gained from this particular study will enable the research team to understand how to generate effective critiques.

We ask that you read this form and ask any questions that you may have before agreeing to be in the study. Anbang Xu is conducting the study under the guidance of Prof. Brian Bailey at the University of Illinois at Urbana-Champaign.

 Procedures

If you agree to be in this part of the study, we would ask you to do the following:

- In the first interview, you will describe your experiences of receiving critiques. Email us three of your designs and the corresponding critiques you received. The digital copies of design artifacts and/or critiques will only be taken if you grant permission. Allow us to inspect these designs and the corresponding critiques. Critiques on the designs will be generated by workers on a crowdsourcing site (Amazon Mechanical Turk). All the designs will be sent de-identified to MTurk workers.

- In the second interview, you will compare critiques generated by the crowdsourcing-based approach with the critiques you previously solicited on you own in terms of the breath, depth, and accuracy of critique.

- Audio recording is optional for the interview. If you are not comfortable with audio recording then just written notes will be taken.

- Consider allowing us to take an electronic copy of your designs for our research paper.
The interviews may be audio recorded to serve as a memory. However, once the interviews are complete, the recording will be transcribed and any personally identifying information will be removed. An interview will last less than 60 minutes on a single day.

Selection Criteria

To participate in this study, we ask that you satisfy the criteria:

- You must be at least 18 years of age.
- You must have experience of receiving design critiques.
- You must be willing to send us three of your designs and the corresponding critiques you received.

Remuneration

You will receive the $30 as long as you attend both of the interviews. If you only attend one interview you will receive $15.

Risks and Benefits of Being in the Study

Subjects will not experience risk beyond that of everyday life. The benefit of being in this study is to contribute further understanding of the nature of design critique.

Confidentiality

The physical and electronic records of this study will be kept private. All the collected designs and critiques will be de-identified when being reviewed by recruited experts. The designs will be destroyed immediately after the critiques are generated by our approach and recruited experts. The audio recording will be destroyed, once the interview has been transcribed. Subjects’ names are not recorded with their responses. In any report that we might publish, we
will not include any information that will make it possible to identify a subject. Records will be kept in a locked file in a locked office or on a secure computer system; only researchers named on this form will have access to the records.

**Voluntary Nature of the Study**

Your decision whether or not to participate in this part of the study will not affect your grades, academic standing or any other relations with the University of Illinois. If you decide to participate, you are free to withdraw from the study at any time. You may skip any question that you do not want to answer.

**Contacts and Questions**

The researchers conducting this study are Prof. Brian Bailey and Anbang Xu from the Department of Computer Science at the University of Illinois at Urbana-Champaign. You may ask any questions that you have now. If you have questions later, you may contact Prof. Bailey, the investigator responsible for this part of the project, by phone at 217-333-6106 or by e-mail at bpbailey@illinois.edu. If you have any questions about your rights as a participant in this study or any concerns or complaints, please contact the University of Illinois Institutional Review Board at 217-333-2670 (collect calls will be accepted if you identify yourself as a research participant) or via email at irb@illinois.edu.

**You will be given a copy of this form to keep for your records.**
Statement of Consent

I have read the above information and questions have been answered. I am 18 years of age or older.

I grant permission for taking copies of my designs and critiques I received    ___Yes  ___ No

• For phone interviews

Do you agree to participate (state ‘yes’ or ‘no’) ?

Are you willing to have the interview audio recorded? (state ‘yes’ or ‘no’) ?

• For in-person interviews

I am willing to participate in the study   ___Yes  ___ No

My two interview sessions can be audio recorded   ___Yes  ___ No

Printed
Name__________________________________________ UIN ________________

Signature __________________________________________ Date______________

Signature of Investigator __________________________ Date______________
Appendix B: Formative Study Interview

About a Person

• Could you describe your educational and/or professional experience in design?

Receiving and Interpreting Critique

• Could you describe a recent or ongoing design project for which you have received one or more critiques? How did these critiques influence the design or your approach to the design?

• How often do you receive critiques for a project? When are critiques most critical in your design process (e.g. parallel design phase or iterative design phase)?

• How would you characterize an effective critique for your work? Can you give specific examples?

• What makes interpreting a critique difficult, if anything?

Types of Critique

• How do you typically receive critiques (e.g. face-to-face, from peers, from experts, or
from online communities)?

• What are the strengths and weaknesses of these methods?

• How often do you use online communities (such as Core 77) for receiving design critiques? Since you can get critiques from your peers, why you still need critiques from online communities? What specifically do you see as the strengths and weaknesses of this method?

Writing Critique and Process of Critique

• How often do you critique the design work by others?

• What are the most important things to communicate about a design or its process when you deliver a critique?

What if a large audience can be instantly reached in your design process.

Image you can instantly reach to a large audience in your design process. This large audience consists of non-experts with an interest in discussing your design.

• What kind of questions do you want to ask the audience? What kind of information do you want to get?
• How this large audience could help you create or improve your design ideas?

Last, would you be willing to share any design artifacts along with the corresponding critiques that were received? These artifacts will remain anonymous, but we may use them for research purposes. For example, we may submit the design artifacts to an online community or through the use of a crowd application to see what types of critiques are generated and compare the critiques to what you originally received.
Appendix C: Lab Study Interview

Introduction to the background of Voyant

We are evaluating a prototype system to help you receive feedback from a non-expert crowd. We would like to learn three things from this discussion: (1) What is the perceived utility of the feedback from your perspective (2) how easy is it to navigate and understand the different types of feedback provided and (3) How can we make the feedback and interaction more useful.

Before you use the system, could you BRIEFLY describe some of the previous feedback you received on this design?

Here is the context information about the system. About 200 people have provided feedback on this design. Most of them don’t have design experience. They are from different counties such as United States, India, UK, and Canada. When you use the interface to explore the feedback, please think aloud, so we can get a sense for what you are seeing and learning. Please also feel free to ask questions.

Perceived quality

- From inspecting all of the feedback generated, what did you learn about your design from the feedback? How does the feedback compare to your expectations? (e.g. did you see anything that was surprising or unexpected?)

- How would you use this feedback to modify or improve the design in another iteration?

- Which type of feedback is most useful for this design?
• Which type of feedback is least helpful? What was missing from the generated feedback, if anything?

• How could you compare the feedback in our system with the feedback you previously received (e.g. feedback from your professors, your colleagues, and online communities)?

• Overall, how would you characterize the quality of the feedback generated?

Influence of the feedback on the design and design process

Assuming we had a fully functional implementation, can you describe when and how often you might use this type of system in your design process? Do you think the availability of this type of system would prompt you to request feedback more often than you do today?

Impact of cost and time

Each type of feedback costs about 3 dollars and takes about 12 hours to generate. Can you speculate how the cost and/or time requirement might affect when or how you might use the system?

How to improve the system

What parts of this interface are particularly useful to you?

What parts of the interface are confusing or otherwise need improvement?
Appendix D: Lab Study Survey

The purpose of this survey is to understand your perception of the feedback generated by our system and its interface. Please cross the circle which best matches your perception. This survey is anonymous (No one will be able to identify you).

<table>
<thead>
<tr>
<th>Please indicate how much you disagree or agree with the following statements:</th>
<th>1 (strongly disagree)</th>
<th>2 (disagree)</th>
<th>3 (slightly disagree)</th>
<th>4 (neutral)</th>
<th>5 (slightly agree)</th>
<th>6 (agree)</th>
<th>7 (strongly agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The “Elements” feedback was helpful.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>2. The “First Notices” feedback was helpful.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>3. The “Impressions” feedback was helpful.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>4. The “goals” feedback was helpful.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>5. The “guidelines” feedback was helpful.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>6. Overall, the generated feedback would help me understand how to improve my design.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>7. I was able to learn something new about my design from the generated feedback.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>8. The system was easy to learn and use.</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>9. The overview of the data</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
presented for each type of feedback was helpful.

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. I was able to easily inspect a subset of the overview data for each type of feedback</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>11. The use of the visual markers is helpful for connecting the feedback to the design</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>12. I would use a system similar to this one to request feedback more often on my designs.</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>
Appendix E: Classroom Study Survey

You have received this survey because you tried the Voyant system (www.crowdfeedback.me). Our goal is the assess ways in which this generated feedback may be helpful to your design and identify areas for improvement.

Please indicate how well the following statements describe your experience using the system:

<table>
<thead>
<tr>
<th></th>
<th>1 (strongly disagree)</th>
<th>2 (disagree)</th>
<th>3 (slightly disagree)</th>
<th>4 (neutral)</th>
<th>5 (slightly agree)</th>
<th>6 (agree)</th>
<th>7 (strongly agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. I was able to learn something new about my design from the feedback.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. The feedback helped me improve my design.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. I am confident in the quality of my design.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. The feedback helped me make cosmetic (surface-level) changes to my design.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
11. Can you describe specific changes made to your design based on the crowd-based feedback received? For each change identified, please also explain your rationale and indicate how significant the change was.

<table>
<thead>
<tr>
<th>Brief description of the change</th>
<th>Rationale (e.g. what feedback prompted you make the change)</th>
<th>How “significant” was the change</th>
</tr>
</thead>
<tbody>
<tr>
<td>17. The feedback helped me make substantive changes to my design.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. The “Elements” feedback was helpful to me.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. The “First Notices” feedback was helpful to me.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20. The “Impressions” feedback was helpful to me.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. The “goals” feedback was helpful to me.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. The “guidelines” feedback was helpful to me.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

150
<table>
<thead>
<tr>
<th>Brief description of the change</th>
<th>Rationale (e.g. what feedback prompted you make the change)</th>
<th>How “significant” was the change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 Cosmetic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 Significant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Brief description of the change</th>
<th>Rationale (e.g. what feedback prompted you make the change)</th>
<th>How “significant” was the change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 Cosmetic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 Significant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Brief description of the change</th>
<th>Rationale (e.g. what feedback prompted you make the change)</th>
<th>How “significant” was the change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 Cosmetic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 Significant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o</td>
</tr>
</tbody>
</table>
12. Can you describe any aspect of the feedback that was particularly helpful?

13. Can you describe any aspect of the feedback that was particularly unhelpful?
Appendix F: Examples for Generating Guidelines Feedback

Task for Generating Feedback on Contrast

Guideline: If the elements (color, size, line thickness, shape, space, etc.) are not the same, then make them very different.

Negative example:

Positive example: Some contrast already happening on the above design, but we can push it further by adding contrast to the elements along the top row.
Task for Generating Feedback on Alignment

Read the design guideline carefully. Compare negative and positive examples of this guideline.

Guideline: Elements of a design should be aligned. You should be able to see an invisible line connects them.

Negative example: This centered alignment appears a bit weak. You can only see a "soft" edge in the design.
Positive example: You can see a "hard" edge on the right in the design.

The invisible line runs right down here, connecting the separate pieces of text.
Task for Generating Feedback on Repetition

Read the design guideline carefully. Compare negative and positive examples of this guideline.

Guideline: Repeating colors, shapes, textures, fonts, spatial relationships, etc. can often help unify the entire design.

Negative example:
Positive example: Repeating the font style more obvious can increase the visual organization and the consistency of the design.

Task for Generating Feedback on Proximity

Read the design guideline carefully. Compare negative and positive examples of this guideline.

Guideline: A design should closely group elements that have relationships. The design should use space to clearly define groups of elements.
**Negative example:** None of the elements connect with any other element in the design.

**Positive example:** This design groups related elements together.


Graham, E.M. (2003). Studio Design Critique: Students and Faculty Expectations and Reality Louisiana State University, Baton Rouge, Louisiana, USA.


165


