

© 2022 Patrick Crain

HOW INTERFACE DESIGN AFFECTS THE COMPOSITION, INTERPRETATION,
AND UTILIZATION OF FEEDBACK

BY

PATRICK CRAIN

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 Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Brian Bailey, Chair
Professor Yun Huang
Professor Alex Kirlik
Dr. Joy Kim, Adobe Creative Intelligence Lab

Abstract

Interfaces for composing and reviewing feedback vary widely in terms of the information they include and how they present it. These interfaces can have a strong impact on how content creators and their feedback providers engage with the feedback, ranging from task performance and perceptions of fairness and helpfulness to strategies used for composing and navigating the feedback. In my thesis, I report the results from three experiments exploring how user interfaces present throughout the feedback exchange process influence feedback composition, interpretation, and utilization. In the first experiment, we examine how various combinations of constructive and summative feedback influence a creator's perceptions of the feedback and the revisions they subsequently perform. Participants (N=441) wrote and revised short stories when presented with quality scores, pre-authored constructive comments, both scores and comments, or no feedback. We found that while showing scores can have marginal benefits compared to showing no feedback, constructive comments without scores led to the most favorable feedback perceptions and revisions. In the second experiment, we investigate how feedback's level of detail influences perceptions of the feedback by both the provider and recipient, and how these perceptions translate to revisions and creative outcomes. Writers (N=285) received feedback from expert providers (N=4) in the form of a writing rubric, open comments, a rubric with open comments, or a rubric with per-criterion comments. We found that writers' revision quality and feedback perceptions increased with the feedback's level of detail, but providers felt the interface that required the most detail diminished their ability to effectively articulate their thoughts. In the third experiment, we explore how students integrate an interactive visualization of feedback's topic and opinion structure into their processes for navigating and interpreting feedback. Teams (N=18) of 3-5 students each used the tool to review feedback on three different project deliverables and revise these deliverables for a grade in an authentic UI design course. We found that students used the visualization to assess their work quality, prioritize revisions, and justify design decisions to their teammates. Some students additionally developed emergent techniques for reviewing and annotating feedback that they adapted to their other projects outside the course. My dissertation represents a large step towards a future where the interfaces with which feedback is composed and presented are given as much consideration as the feedback's content when designing tools for supporting feedback exchange between content creators and feedback providers.

Acknowledgments

The last 7 years have been far and away the most challenging years of my life for a variety of reasons, and I could not imagine having the strength to pursue – much less complete – my Ph.D. without the love and support from dozens of people rooting for me to succeed.

I would like to first and foremost thank my advisor, Dr. Brian Bailey, for not only providing constant support and guidance, but having an unreal amount of patience in mentoring me as I repeatedly stumbled in learning how to conduct and report research of the highest quality. To this day it is difficult to imagine what kind of potential you saw in the inexperienced, incoherent, and socially awkward person I was when I began my journey at UIUC. From helping me write half of my first published paper to giving me just enough support to write my final papers on my own, I am sincerely grateful for the several chances you took in giving me the time, resources, and opportunities I needed to develop and hone my skills as an independent researcher.

I would next like to thank the distinguished researchers who served on my committee and provided additional mentoring support throughout the years. To Dr. Yun Huang, I am grateful for your inviting me to attend your own group's meetings to brainstorm research ideas with your students and yourself. I thoroughly enjoyed the few discussions I was able to attend, and several topics from these meetings sparked ideas that made their way into my last few papers in significant ways. To Dr. Joy Kim, I am grateful for your ability to present alternative perspectives with a positive and constructive energy. Your perspectives have not only been useful in improving the quality of my research, but have served as wonderful examples I strive to mirror in presenting my own ideas to others. To Dr. Alex Kirlik, I am grateful for your taking the time to offer feedback and insights on my experimental designs. Although we sat down to chat maybe once or twice per year, each time left me feeling much more confident in my experiments, and at least two of my papers are vastly improved as a result of our discussions.

I would also like to thank my family for their unwavering support not just throughout graduate school, but throughout the entirety of my life and education. To my mom and dad especially, thank you for being incomprehensibly invested in every aspect of my happiness and well-being. Regardless of what my needs are or what my pursuit of happiness may entail, and regardless of how busy, cranky, or stressed I get, both of you have always been there to help me in every way I could possibly imagine and in many ways I couldn't imagine. I am immensely grateful to have such supportive and loving parents, and I hope to never take that

for granted. To my sister, thank you for loving and caring for me in the goofball way that you do. Even when I'm heads down working for weeks on end, your excitement in sending me cute pictures from the vet's office never fails to bring a chuckle and a smile to my face when I finally remember to check messages once in a while. To my grandma, thank you for continuing to think of me and keeping me part of your life even from thousands of miles away. I love you all very much, and I am truly blessed to call you my family.

I would of course like to thank all of my wonderful friends for their support and companionship throughout our mutual journeys, especially Luke Upton, Sneha Kumaran, and Peyton Smith. Luke, you've been one of my closest friends for 20 years now, and getting to share and bond over parallel life experiences for two decades is a truly amazing feeling. Whether it's playing Command and Conquer, commiserating over research duties, sharing random videos, or contemplating more serious topics, hanging out with you is always a bright spot in my day, and I am excited you will soon complete your Ph.D. journey as well. Sneha, you are the first and closest friend I've made at UIUC, and watching you succeed as we both fought similar battles was probably the biggest motivator for me to push myself for the majority of graduate school. Beyond that, you've been a wonderful friend and have helped with my research and my life in several ways I don't think I can repay, and I wish you the best with your new family and job. Peyton, you are one of the wisest, most understanding, and most sincere people I have ever met, and whether I'm practicing Super Smash Brothers Melee with you or picking your brain about philosophy, almost every time we hang out is an eye-opening experience. Even if they're months apart, I always look forward to each time we get to talk and hang out, and I hope you are able to find happiness and fulfillment in all of the things you do. I would also like to thank Grace Yen, Sebastian Rodriguez, Charlotte Yoder, Wayne Wu, Gina Do, Emily Hastings, Helen Wauck, Wendy Shi, Kristen Vaccaro, Robert Deloatch, John Lee, and all of my other friends and lab mates at UIUC for the company and moral support that have made the last 7 years that much easier.

Finally, I would like to thank my extraordinary girlfriend and future wife, Kaii Ki, for continuously supporting me in countless ways each and every day. Words cannot describe how much of my success, happiness, and well-being the last few years is the result of your love, understanding, patience, wisdom, and strength. Between keeping me company as I pull all-nighters trying to revise papers, making sure I get the nourishment I need as I push myself each day, and doing everything you can to brighten my mood when I'm stressed or cranky, I could not imagine a more loving or caring person to spend the rest of my days with.

Thank you all!

Table of Contents

Chapter 1	Introduction	1
1.1	Vision	3
1.2	Prior Work	3
1.3	My Work	5
1.4	Scope	9
1.5	Use Case Scenario	11
1.6	Contributions	13
Chapter 2	Related Work	16
2.1	Helping Creators Interpret Feedback	16
2.2	Helping Feedback Providers Meet Creators' Needs	20
Chapter 3	Preliminary Work: Exploring How Designers Approach Iteration in an Online Critique Community	26
3.1	Introduction	26
3.2	Research Questions	28
3.3	Methodology	28
3.4	Results	31
3.5	Discussion	42
3.6	Limitations and Future Work	44
3.7	Contributions	45
Chapter 4	Determining How Scores Mediate Interpretation of Written Comments	47
4.1	Introduction	47
4.2	Research Question and Hypotheses	49
4.3	Methodology	49
4.4	Results	56
4.5	Discussion and Future Work	62
4.6	Limitations	64
4.7	Contributions	65
Chapter 5	Investigating How Feedback Detail Affects Feedback Composition and Interpretation	67
5.1	Introduction	67
5.2	Research Questions and Hypotheses	69
5.3	Methodology	70
5.4	Results	78
5.5	Discussion	83

5.6	Limitations and Future Work	85
5.7	Contributions	86
Chapter 6 Exploring How Visualizing Feedback’s Topic and Opinion Structure		
	Influences Feedback Exchange	88
6.1	Introduction	88
6.2	Research Questions	91
6.3	Feedback Visualization Tool Studied	91
6.4	Methodology	93
6.5	Results	102
6.6	Discussion	116
6.7	Limitations and Future Work	120
6.8	Contributions	121
Chapter 7 General Discussion		
7.1	Personalizing Feedback’s Presentation	123
7.2	Generating and Presenting Metadata	125
7.3	Facilitating Composition in Different Contexts	128
7.4	Making Feedback Exchange More Accessible	129
7.5	Generalizing Insights to Different Feedback Exchange Contexts	130
Chapter 8 Conclusion		
References		134

Chapter 1: Introduction

This dissertation contributes new empirical data, design knowledge, and theoretical knowledge regarding how user interfaces impact content creators and their feedback providers throughout the feedback exchange process. Throughout this dissertation, I use the term “content creator” (or just “creator”) to refer to anyone who employs open-ended processes to produce digital content or intellectual property. Such creators might include someone who writes a short story using a word processor, prototypes a user interface using a digital mockup tool such as Balsamiq, or designs a research poster using an image editing tool such as Photoshop.

Creators rely extensively on feedback exchange to iteratively refine their work, develop their skills, and deepen their understanding of their craft [1]. Effective feedback exchange requires a joint effort between creators and their feedback providers, with unique challenges on either end. Providers need to communicate their ideas and insights in a way that is meaningful to the recipient, while a creator must in turn interpret and extract meaning from the feedback they receive as it relates to their project design goals. A creator’s feedback needs may also change throughout their design process (e.g., conceptual feedback for early drafts vs. concrete suggestions for later ones), potentially making some forms of communication more effective than others at different design stages. For both feedback composition and interpretation, user interfaces play a central role in facilitating effective communication throughout online feedback exchange (see Figure 1.1).

Existing works investigating how interfaces impact feedback exchange have three main limitations. First, although many interfaces supplement constructive comments with scores, prior works exploring how scores impact feedback interpretation report mixed results. Prior works show that scores can signal room for improvement [2], motivate revision [3], and maintain task interest [4], but can also downplay the need to improve [2] and cause feelings of discouragement, self-doubt, and inadequacy [4, 5]. These contrasting findings indicate that additional guidance in using scores may be necessary. Second, prior work often considers the perspective of only the feedback provider or only the feedback recipient when assessing an interface’s effectiveness, ignoring potential tradeoffs from either perspective. For example, simplifying feedback composition may diminish the recipient’s perceived quality and usefulness of the feedback, while providing highly detailed feedback may not be worth the cost and difficulty of composing feedback at such detail. Third, prior work does not examine how interfaces facilitate and extend the unique tasks creators pursue when interpreting unstructured feedback in practice, such as prioritizing and identifying patterns in

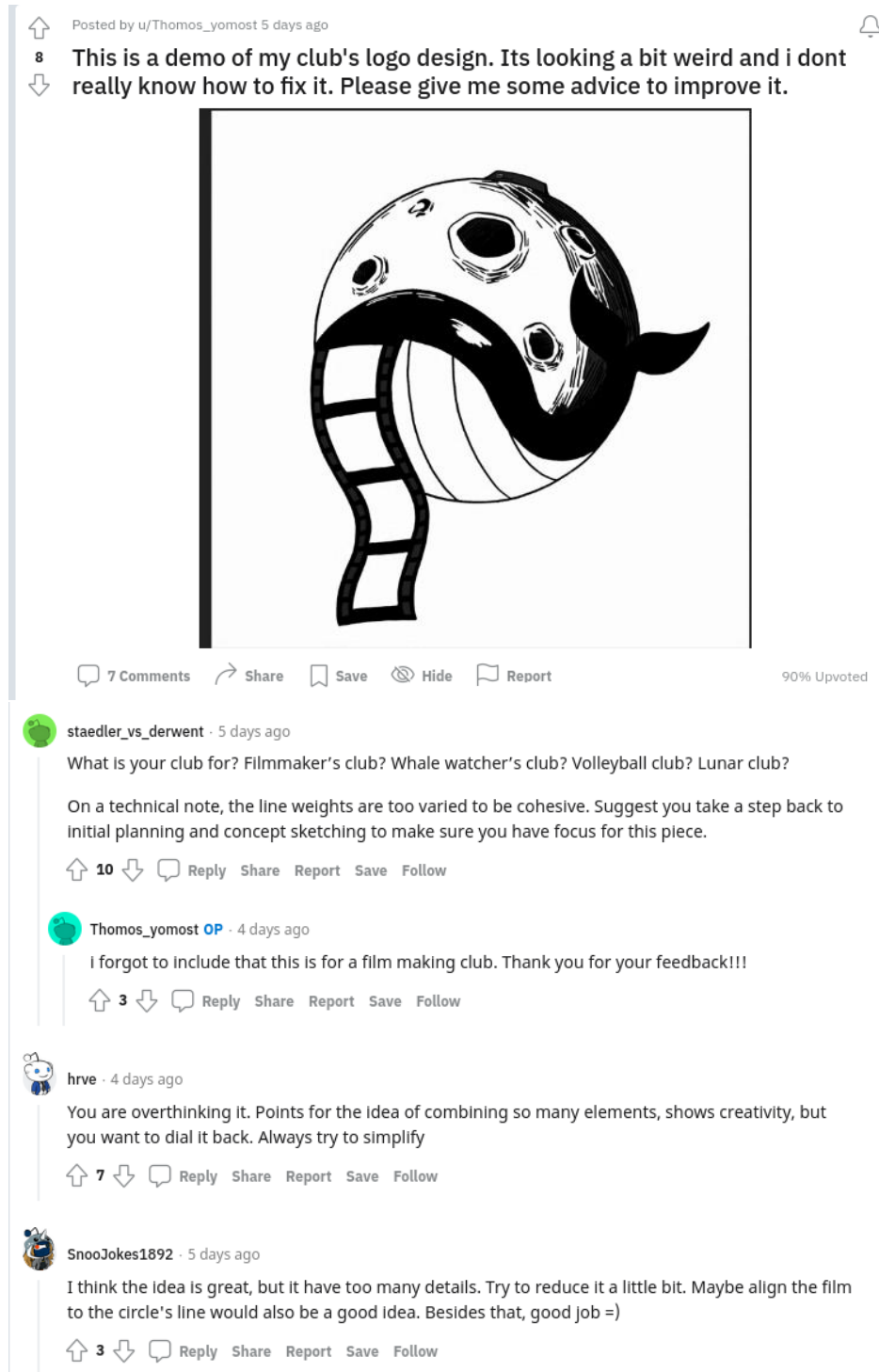


Figure 1.1: An example of a graphic design (top) posted to the /r/design_critiques sub-Reddit and comments (bottom) providing feedback on the design. The study described in Chapter 3 revealed that the organization of feedback into reply threads and the ability to anonymously engage with a diverse range of feedback providers are key user interface features that make the community especially appealing to novice designers. These findings motivated the further exploration of how user interfaces influence communication between creators and their feedback providers throughout the remainder of this dissertation.

the feedback. Understanding how user interfaces influence both creators and their feedback providers throughout the feedback exchange process is critical to facilitating communication and helping creators achieve their design goals.

My dissertation takes concrete steps towards addressing the above limitations by investigating 1) how the technique and interface used to compose feedback influence a provider’s composition process and perceptions of the feedback they compose, 2) how the interface and information with which feedback is presented influence a creator’s perceptions and interpretation of the feedback they receive, and 3) how interfaces for both composing and presenting feedback ultimately influence the depth, type, and content of revisions creators perform in response to the feedback. I envision a future where the interfaces with which feedback is composed and interpreted are as thoughtfully considered as the content of the feedback itself.

1.1 VISION

My vision of the future is one where creators can readily solicit feedback personalized in both its content and its presentation, and where feedback providers can organically offer feedback based on the recipient’s needs and their own abilities without inhibition by choice of composition interface. Such a future will allow for more effective communication between creators and their feedback providers throughout the feedback exchange process. This communication in turn will result in feedback that is higher in quality, that is easier to interpret, and that leads to more effective revisions. This dissertation progresses towards this vision by investigating how several aspects of interfaces for composing and reviewing feedback such as their mode of evaluation (summative vs. constructive), level of detail, and visual representation influence the feedback exchange process.

1.2 PRIOR WORK

The HCI community has historically used interface design to improve the quality of feedback composition using one of two approaches. The first approach has been to scaffold an individual feedback provider’s composition through interventions such as rubrics [13, 23] or comparisons [6, 11]. The second approach has been to break down feedback composition into microtasks and delegate these microtasks to crowd workers [7, 8, 9, 24]. Although each of these approaches has its own merits, most research investigating either approach considers only the feedback recipient’s perceptions of the feedback’s quality. As a result, they largely

		Stage of Feedback Exchange Examined (R) = Recipient, (P) = Provider										Interface Design Choices Contrasted					Experimental Design Used				
		(R) Seeking Feedback	(R) Soliciting Feedback	(P) Planning Critique	(P) Critiquing	(P) Assessing Critique	(R/P) Assessing Feedback	(R) Interpreting Feedback	(R) Implementing Feedback	(R) Revision Outcomes	(R) Learning Outcomes	Framing	Structure	Cues	Medium	Interaction	Quantitative	Qualitative	Formative	Field Study	Controlled
This Thesis	Chapter 3						x	x	x	x			x	x			x	x		x	
	Chapter 4						x	x	x	x		x		x			x				x
	Chapter 5				x	x	x	x	x	x		x	x				x	x			x
	Chapter 6		x	x	x		x	x	x	x				x	x	x	x	x		x	
Composition Interventions	Kang et al. [6]			x	x		x										x	x	x		
	Xu et al., 2014 [7]						x	x									x	x	x		
	Greenberge et al. [8]						x	x									x	x	x		
	Luther et al. [9]						x	x									x	x	x		
	Cheng et al. [10]		x	x	x	x						x	x				x	x		x	
	Cambre et al. [11]				x	x	x							x			x				x
	Hicks et al. [12]						x					x	x				x				x
	Yuan et al. [13]						x	x				x					x				x
Interpretation Interventions	Schneider et al. [14]		x				x					x					x				x
	Xu et al., 2015 [15]						x	x		x		x		x			x	x		x	
	Yen et al., 2017 [16]							x	x	x						x	x	x			x
	Wu et al., 2018 [17]							x	x	x		x					x				x
	Butler et al. [18]							x		x				x			x				x
	Lipnevich et al. [19]						x	x		x		x		x			x				x
	Wu et al., 2016 [20]						x	x						x			x				x
	Wu et al., 2017 [21]						x	x		x		x					x				x
Interpretation Interventions	Yen et al., 2020 [22]						x								x		x				x

Table 1.1: This table summarizes prior work that examines how interface design choices may influence the feedback exchange process. The table shows which steps of the feedback exchange process each work covers, which interface design choices are contrasted, and which types of experimental designs are used. Works covered in this dissertation are indicated in **boldface**.

ignore the feedback providers’ perspectives on the difficulty and effectiveness of composing the feedback, both of which are critical concerns when designing in practice.

Prior work has demonstrated that even when feedback is of high quality, creators may still struggle to interpret it [16, 25, 26, 27]. The HCI community has consequently developed and tested several interfaces for helping creators interpret the feedback they receive. Some of these interfaces promote self-regulated behaviors such as reflection to help creators engage with their feedback [16, 28] or enhance their resilience to harsh criticism [17]. Others aim to frame the feedback through means such as including praise [18, 19] or presenting cues about the feedback provider’s expertise [20]. Another class of interfaces leverages visualization

techniques to help creators explore feedback metadata, such as using visual markers to identify areas of a graphic design referenced in the feedback [15]. Assessments of these interfaces often do not examine how creators interpret feedback in practice, highlighting a gap in knowledge regarding the unique processes enabled by different interfaces that creators develop and pursue to achieve their design goals.

Interfaces influence the entire feedback exchange process, which begins with a creator’s initial decision to seek feedback and ends with the incorporation of that feedback into a revised design. The experimental designs of existing works typically examine how a given interface impacts feedback exchange at a few important but isolated stages. Although there are often logistical and practical reasons for such examinations, determining how interface choices impact other steps of the feedback exchange process can be challenging without speculation. For example, it is not always apparent whether simplifying feedback composition makes feedback interpretation more difficult, or vice versa. My dissertation takes steps towards addressing the above challenges by examining how user interfaces affect both providers and recipients throughout the feedback exchange process.

1.3 MY WORK

My dissertation contributes to online feedback exchange by exploring how the interfaces used for composing and interpreting feedback influence the feedback exchange process. To this end, I conduct three interconnected research studies investigating how presenting constructive feedback with summative feedback (Chapter 4), at different levels of detail (Chapter 5), and through a visualization (Chapter 6) influences the composition, interpretation, and utilization of feedback. In this section, I describe the research methodologies for each of these studies in greater detail.


Presenting Constructive Feedback with Summative Feedback: Quality scores are used in several domains and disciplines to supplement constructive feedback, but are underexplored in their effectiveness for facilitating revisions to open-ended creative works. In Chapter 4, I investigate how novice writers ($n=441$) compose and revise short stories with respect to either pre-authored constructive written comments, a numeric quality score, both comments and a score, or no feedback (see Figure 1.2). The factors and study domain were chosen to mirror real-life automated feedback systems on open-ended work, wherein quality scores often accompany pre-authored constructive written comments. The goal of this study is to determine how showing a score can shift creators’ perceptions of the feedback they receive, how these perception influence the subsequent revisions they make to their work, and whether showing scores alongside written comments is ultimately beneficial or detrimental

Write a Short Story

Instructions

Please spend 10-15 minutes writing a short (125-250 word) story based on the writing prompt below. The story must be your own original work; please do not reuse or copy from an existing story. The task requires that you compose your story in the text box below. Please do **NOT** write the story in a different software tool and then paste the content; this action may disqualify the submission and payment. Your work will be autosaved every few seconds.

Prompt



Write a short story based on the image above.

Please write your story in the space below:

Harry let out a long sigh, turning back to look into the suitcase again. A small, pink unicorn toy, about a quarter of Harry's size, was pacing back and forth on the bottom of the suitcase, intermittently trying to leap out. Harry leaned down, scooped the unicorn up in one paw, and deposited him in the dirt beside the suitcase.

"Come on, man!" the unicorn explained in a deep tenor. "Can you believe this happened again, man? How many times is this going to happen? First Frankie ditches us, then Tommy, and

Current word count: **247 words**

[Submit Story](#)


Revise Your Story

Welcome back! Your story has been reviewed by the judge panel and has received the following evaluation:

Score: 75/100

Feedback: The plot and setting are reasonably well-established; however, the story can be further improved by including more details about the plot or setting in which the story takes place, or providing more context for the opening or ending of the story. The characters are somewhat simplistic and under-developed, so revealing more about them by communicating their thoughts, appearance, action, or dialogue would better captivate readers. The story makes little use of physical and visual language, so there is plenty of room to enhance the story's quality by incorporating more lively language that appeals to the senses and vivid descriptions of the characters and their environment. Please note that while each of the components are graded separately, they work together, and that directed improvement toward one area may indirectly improve another. This feedback is meant to give you an idea of the areas in most need of general improvement.

Prompt



Write a short story based on the image above.

Figure 1.2: Many automated assessment platforms combine quality scores with preauthored comments to facilitate feedback exchange at scale. However, prior work reports mixed results concerning the value of presenting scores on creative works. I conducted a study (Chapter 4) where writers were asked to draft a short story (left image above) and revise it with respect to feedback presented through one of four interfaces. These interfaces presented feedback as either a quality score, constructive written comments, both a score and comments (right image above), or a prompt without feedback (i.e., “Your story has been reviewed by the judge panel and is ready for revision”). This study measured writers’ perceived feedback helpfulness and fairness, the types of revisions they performed, the effort they put into these revisions, and the quality of their initial and revised story drafts.

to the interpretation and operationalization of those comments. To this end, I use post-task survey responses and data collected during the story-writing task to measure participants’ perceived helpfulness and fairness of the feedback they receive, as well as the effort, depth, and quality of the revisions they perform in response to that feedback.

Criteria	Excellent	Above Average	Developing	Needs Improvement
Narration	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to a maximum level .	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to an acceptable level .	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to a minimum level .	World of story is presented as inconsistent and unbelievable through no or improper use of details.
Characterization	Characters are often revealed indirectly through their physical appearance, action, thought, dialogue, setting, and symbol.	Characters are sometimes revealed indirectly through their physical appearance, action, thought, dialogue, setting, and symbol.	Readers are often directly told what a character is like with little explanation of characters' thoughts and actions, or use of dialogues, etc. to reveal the characters.	Readers are directly told what a character is like with no explanation of characters' thoughts and actions, or use of dialogues, etc. to reveal the characters.
Imagery	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to a maximum level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to an acceptable level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to a minimum level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are not used in the story.
Mechanics	No grammatical, punctuation, and spelling errors exist in the work.	Minimal grammatical, punctuation, and spelling errors exist in the work.	Some grammatical, punctuation, and spelling errors exist in the work.	Many grammatical, punctuation, and spelling errors exist in the work.
Structure	There is wide variety in sentence structure.	There is some variety in sentence structure.	There is minimal variety in sentence structure.	There is no variety in sentence structure.
Reviewer's Comments	This story fell flat for me in a few ways. The world didn't feel believable to me because I felt we didn't receive any description of or rooting in it. The biggest problem, in my opinion, is the characterization problem. To a degree, it felt like there might as well not be any characters in this story. I want to see their interactions, meaningful ones, not simply be told what their interactions are like.			
Save Feedback			Flag for Admin Review	

Revise Your Short Story

Welcome back! Another participant in this study has evaluated your story. The evaluation of the story according to the criteria in the rubric are **highlighted in blue**. Any additional overall comments are shown in the last row of the table.

Instructions

The story that you previously submitted has been copied into the text box at the bottom of the page. Please **review the feedback below** and **spend 5-10 minutes revising your story** with respect to the feedback and make any additional revisions you'd like. Please press the submit button when you are finished. The word limit has been increased to 300 words. Do **NOT** revise the story in a different software tool and paste the content into the text box below. This action may disqualify the submission and payment. Your work will be autosaved every 10 seconds.

Feedback

Criteria	Excellent	Above Average	Developing	Needs Improvement
Narration	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to a maximum level .	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to an acceptable level .	Specific details that contribute to the creation of an authentic, consistent, and believable world of story are used to a minimum level .	World of story is presented as inconsistent and unbelievable through no or improper use of details.
Characterization	Characters are often revealed indirectly through their physical appearance, action, thought, dialogue, setting, and symbol.	Characters are sometimes revealed indirectly through their physical appearance, action, thought, dialogue, setting, and symbol.	Readers are often directly told what a character is like with little explanation of characters' thoughts and actions, or use of dialogues, etc. to reveal the characters.	Readers are directly told what a character is like with no explanation of characters' thoughts and actions, or use of dialogues, etc. to reveal the characters.
Imagery	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to a maximum level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to an acceptable level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are used to a minimum level in the story.	Concrete and significant details that appeal to senses and suggest ideas beyond the surface are not used in the story.
Mechanics	No grammatical, punctuation, and spelling errors exist in the work.	Minimal grammatical, punctuation, and spelling errors exist in the work.	Some grammatical, punctuation, and spelling errors exist in the work.	Many grammatical, punctuation, and spelling errors exist in the work.
Structure	There is wide variety in sentence structure.	There is some variety in sentence structure.	There is minimal variety in sentence structure.	There is no variety in sentence structure.
Reviewer's Comments	This story fell flat for me in a few ways. The world didn't feel believable to me because I felt we didn't receive any description of or rooting in it. The biggest problem, in my opinion, is the characterization problem. To a degree, it felt like there might as well not be any characters in this story. I want to see their interactions, meaningful ones, not simply be told what their interactions are like.			

Figure 1.3: Feedback composition interfaces vary in the time, effort, and skill sets feedback providers require to communicate their thoughts to creators. However, little work has investigated the tradeoffs between the costs of using different interfaces to compose feedback and the value of that feedback to creators. In Chapter 5, I investigated how feedback providers composed feedback at four different levels of detail, and whether they perceived the difficulty of composing feedback at each level outweighed the value of that feedback. I also investigated creators' perceptions of feedback fairness, personalization, and helpfulness, as well as the type and quality of revisions they performed in response to reviewing feedback at each level. The images above depict the rubric + open comments interfaces for composing (top) and reviewing (bottom) feedback. The presentation of feedback in each composition interface was chosen to mirror the presentation of feedback in its corresponding review interface.

Presenting Constructive Feedback at Different Levels of Detail: The complexity of feedback composition varies widely depending on the interface used for the task, ranging from clicking a few rubric items to writing free-form comments in response to several prompts. However, little work has directly explored the tradeoffs of feedback constructed at different levels of detail as perceived by both creators and their feedback providers. In Chapter 5, I investigate how presenting feedback at four different levels of detail with increasing specificity and elaboration impacts participants' ($n=285$) perceptions of helpfulness and revision quality during a creative short story writing task. I also investigate how the increased time and effort required to compose increasingly detailed feedback impacts providers' ($n=4$) perceptions of the feedback's utility. The levels of detail from least to most specific were 1) rubric-derived comments, 2) free-form personalized comments, 3) rubrics with personalized free-form comments (shown in Figure 1.3), 4) and rubrics with personalized comments on each rubric item. These levels were chosen as representatives of the types of feedback presentations currently prevalent among online feedback exchange platforms and communities. One goal of this study is to explore how presenting feedback at different levels of detail influences creators' perceptions of feedback helpfulness, fairness, and credibility, as well the depth and quality of the revisions they make. Another goal is to determine whether the (presumably) increased difficulty of composing feedback at higher levels of detail is worth the increased benefit (if any) to the feedback recipients. To achieve these goals, I use post-task survey responses and interaction logs to measure writers' self-reported perceptions of feedback helpfulness, feedback fairness, revision effort, and revision depth. Additionally, I use a parallel set of surveys and interaction logs to assess the amount of time and effort providers spend composing feedback at each level of detail, as well as the providers' perceptions of the helpfulness and expressiveness of the feedback they compose.

Presenting Unstructured Written Feedback as a Visualization: Unstructured written feedback can be difficult for creators to interpret, especially in large quantities. While prior work suggests visualizations can help creators reconcile conflicting opinions in unstructured feedback, these works do not explore the mechanisms or processes creators leverage when using visualizations to interpret feedback. In Chapter 6, I conduct a field study exploring how students in an authentic classroom environment use an interactive visualization to review unstructured feedback from several providers with potentially conflicting opinions (see Figure 1.4). The visualization organizes feedback by its topic, opinion, and provider to help students identify similar comments on high level aspects of their designs, as well as individual providers' opinions on these aspects (e.g., whether an aspect was handled well or poorly, whether they have questions about a design decision or suggestions for improving it, etc.) The use of the visualization was integrated into the instruction of a 12-week user

interface design course, wherein 18 teams of 3-5 students each used the visualization to review feedback for three project deliverables and revise them for a grade. This study aims to determine how and whether visualizing the topics and opinions in a collection of feedback enables creators to effectively interpret the feedback they receive. I also aim to identify the tasks creators perform and the goals they pursue that uniquely arise from reviewing a visual representation of the topics and opinions within a collection of feedback (e.g., reviewing clusters of opinion icons to identify weak areas of a design). To measure these items, my exploration draws data from several sources including responses from surveys associated with each project deliverable, interaction logs, and interviews with multiple students (n=12) and teaching assistants (n=2).

1.4 SCOPE

This dissertation explores how interfaces for composing and interpreting feedback influence the process of feedback exchange between creators of iteratively revised projects and their feedback providers. The individual studies described in this dissertation target feedback on a single initial or early draft of a creative project, where creators are generally more open to larger changes to their work and where feedback’s impact is most noticeable. Although insights from individual studies may generalize to additional feedback-revision cycles, this dissertation does not directly explore how user interfaces influence feedback exchange across multiple revisions of the same project.

The studies presented in this dissertation examine how the representation and composition of feedback influence creators and their feedback providers across several stages of the feedback exchange process. These stages extend from the provider’s initial critique planning to the creator’s design outcomes after incorporating the feedback into their revisions. This dissertation does not examine the earliest stages of feedback exchange in which a creator decides where, when, and how to initially seek feedback on their work, though it is possible the tool in Chapter 6 could help creators decide when and from whom to seek feedback for later iterations of the feedback-revision cycle. Examinations of the feedback exchange process beyond revision outcomes (i.e., long-term learning or skill development) are likewise out of scope.

Prior work has explored several feedback presentation interfaces individually in terms of how they affect performance on creative open-ended tasks [7, 9, 12, 29, 30, 31]. This dissertation focuses on comparing these methodologies to one another and exploring how factors other than performance and revision quality are influenced by the interface used to present feedback to creators. This dissertation also does not explore all possible presentations of feedback, but



Figure 1.4: Creators often struggle to interpret collections of unstructured written feedback, especially when it contains diverse or conflicting opinions. While prior work suggests visualization tools can aid feedback interpretation, these works do not explore how such tools facilitate creators' processes for accomplishing tasks unique to interpreting unstructured feedback. Chapter 6 investigates how creators leverage a visualization tool to interpret feedback, as well as the goals and processes they develop for interpreting feedback using the tool. The tool allows creators to navigate feedback by provider and assign topic and opinion labels to each feedback statement (top image). After the feedback is labeled, the tool presents a creator with an interactive visualization (bottom image) which organizes the feedback into a grid by provider (row), topic (column), and opinion (icon shape and color). Clicking an icon allows a creator to assign intent-to-act labels to each piece of feedback, which can later be reviewed using the filter bar on the top.

rather a representative sample of common feedback presentation methodologies sufficient to paint a broad picture of how interfaces can influence perceptions of feedback and subsequent revisions.

Creators who received feedback in the studies described in this dissertation consisted primarily of novices in their respective fields. This dissertation does not explore potential differences in how experts might perceive, interpret, and implement the feedback they receive, although the data presented in this dissertation can provide a baseline for future analyses. Finally, this dissertation examines feedback exchange within the disciplines of graphic design, creative writing, and user interface design. Each of these domains is representative of a class of domains in which creators produce and iteratively refine an open-ended project based on feedback provided by stakeholders (such as instructors, employers, clients, or users). Though not directly within scope, the insights from this dissertation may generalize to other domains such as programming, music composition, or academic writing.

1.5 USE CASE SCENARIO

Imagine Alice, a full-time college freshman majoring in graphic design with a passion for learning her craft. She is in the conceptual stages of designing a logo for a class project, and seeks fresh perspectives on the design from others outside her class. Alice approaches an online community for feedback on her design. She navigates to the community’s “Share” page, where she is invited to upload her design and specify her design stage (early vs. late, Chapter 3), preferred level of feedback detail (high-detail vs. low-detail, Chapter 5), and any other details she feels might help the community provide more appropriate feedback (Chapter 3).

Within a few hours, several community members view and comment on Alice’s design. One such member is Carol, an experienced freelance graphic designer. Carol reviews Alice’s design and, upon clicking a “Critique” button, is presented an open text box per her own writing preferences (choice of composition UI, Chapter 5), and asked to avoid empty praise and criticism per Alice’s feedback preferences (summative feedback, Chapters 4 & 5). Before submitting her feedback, Carol is given the option to share her design background and expertise (provider background, Chapter 6) or to remain anonymous (Chapter 3), choosing the former. This dissertation advances the idea that giving Carol agency over how she composes feedback will help her more effectively communicate her insights to Alice, who will in turn benefit from better feedback and stronger revisions.

When Alice returns, she is presented with her feedback as a list sorted by provider in order of decreasing expertise (organization, Chapter 6) after filtering empty praise and

criticism (hiding summative feedback, Chapter 4), ensuring she can quickly locate insights from experienced designers. As she reviews her feedback, Alice notices Carol has asked her to consider how the logo would look in black and white as a benchmark for its visual integrity at different sizes. Although no other comments suggested this technique, Alice defers to Carol’s expertise and experiments with black and white logo variations. Alice soon realizes she is unable to make her current logo look appealing in black and white, and revises her design until the monochrome version is to her liking. After converting her latest revision back to full color, Alice affirms the improved visual integrity of her logo and notes Carol’s monochrome technique for future reference.

A few days later, Alice returns to the community for a second round of feedback on her most recent logo revisions. Alice uploads her latest revisions to the community, specifying she is now looking for late stage feedback while leaving her other preferences the same. Carol comes across Alice’s revised design and, upon clicking the “Critique” button, is asked to rate various aspects of Alice’s logo (summative feedback, Chapter 4) and to write a few sentences explaining each rating (per-criterion comments, Chapter 5). Carol rates each aspect of Alice’s design, but feels her writing style makes it difficult to separate her comments into individual design aspects (composition style, Chapter 5). Carol uses a drop-down to select an interface for writing comments in a single text box while still allowing her to rate individual design aspects (rubric scores + open comments, Chapters 4 & 5).

Upon returning to view her feedback, Alice is prompted to review her feedback in plain-text form and assign topic labels to each statement to match her organizational style (self-annotation, Chapter 6). After completing the labeling, Alice is presented a topic and opinion visualization that organizes feedback alphabetically by topic (Chapter 6) and displays quality scores (Chapter 4) for each topic based on sentiment analysis. Results from this dissertation suggest visualization can help Alice make sense of feedback without reading it in its entirety, allowing her to prioritize revisions more quickly than by reading plain text feedback. This dissertation’s results also suggest involving Alice in constructing the visualization through labeling can promote a deeper understanding of the feedback, ensuring she is able to effectively identify and address the most important feedback.

As Alice scans the quality scores next to each aspect of her design, a single highlighted score stands out next to the “colorblind friendly” topic. Reading this feedback helps Alice quickly learn that while her logo is legible to those with red-green colorblindness like herself, it is less legible to those with blue-yellow colorblindness. After tweaking the palette of her logo to address the visual accessibility concerns, Alice makes a note to seek out and pay special attention to feedback on visual accessibility for her future projects.

The above scenario illustrates how personalizing the interfaces used to compose and present

feedback can facilitate stronger communication throughout the feedback exchange process. While this scenario depicts one path of how a creator might leverage the results of this dissertation, a creator might use only a subset of these interfaces for a design project in practice, or may use alternative interfaces. For example, a creator might have left the generation of topic metadata to a third party if they had the resources to do so, or may have skipped it altogether if it did not suit their needs. In any of these scenarios or others, this dissertation advances the idea that personalizing the interfaces used throughout the feedback exchange process to meet the unique needs of creators and feedback providers can aid with the feedback interpretation and composition.

1.6 CONTRIBUTIONS

This dissertation contributes theoretical, empirical, and design knowledge regarding how feedback presentation influences the way creators and their feedback providers perceive feedback, and how these perceptions in turn influence creators' revision behavior. My dissertation also extends a feedback visualization tool and contributes practical design recommendations for presenting feedback, which may be leveraged to help creators to achieve their desired revision outcomes while working within their individual constraints. The specific contributions are as follows:

Empirically-derived guidelines for combining summative and constructive feedback on creative works. In Chapter 4, we analyze how the presence of a quality score changes a creator's perceived fairness and helpfulness of pre-authored written feedback on their creative and open-ended work, and how the presence of this score influences their revision effort, depth, and outcomes. Our results suggest that constructive comments on open-ended work generally promote the most positive feedback perceptions and effective revisions when presented without a quality score. However, task performance and self-reported task satisfaction were positively correlated when quality scores were shown, indicating scores may still have value in affirming creators' of their high quality work. These findings supplement evaluations of interfaces for automated assessment systems that leverage quality scores in their feedback output, and serve as a reference point for subsequent studies evaluating alternative feedback presentations. A paper reporting the results of this study was published in ACM Learning at Scale, 2021.

A taxonomy of design recommendations for composing and presenting feedback at different levels of detail. In Chapter 5, we investigate how creators and their feedback providers perceive feedback organized at four different levels of detail. Our main result was that while participants' revision quality and favorable perceptions of the feedback increased

with the feedback’s detail, providers felt better able to communicate their ideas using a free-form comment interface than when using a more detailed interface with comments on each rubric item. We also found that providers spent half as much time composing feedback using rubrics alone than using rubrics with per-item comments, and that rubrics alone still enabled participants to make revisions that improved the overall quality of their work. We distill these findings into a taxonomy of design recommendations for using feedback at different levels of detail, such as using per-item comments to maximize perceived revision quality, open comments to maximize perceived feedback quality, and rubrics to maximize feedback-revision turnaround. A paper reporting the results of this study was published in *ACM Computer Supported Cooperative Work*, 2022.

Results from a field exploration of how visualizing topics and opinions affect feedback engagement. In Chapter 6, we extend an interactive visualization tool for exploring feedback based on high-level metadata such as topic, opinion, and provider information. We deploy this tool in an authentic classroom environment to determine how presenting feedback through an interactive visualization can enhance and extend students’ processes for interpreting multiple pieces of feedback. Our field study revealed that students leveraged the visualization to assess their work’s quality, prioritize revisions based repeated suggestions and criticisms, and justify design decisions to their teammates. We also found that students reported deeper familiarization with their feedback when they labeled it with topic and opinion metadata themselves compared to when it was labeled for them, highlighting benefits of involving students in the visualization process. This work contributes empirical data regarding the effectiveness of supporting feedback interpretation through interactive visualization, as well as several guidelines for leveraging such visualizations effectively within and beyond the classroom. A paper reporting the results of this study is under submission to *ACM Transactions on Computer Human Interaction*, 2022.

Empirical data and guidelines for how creators iterate on their work using feedback from an online community. In Chapter 3, we explore why, when, and how graphic designers approach a public online design community to solicit feedback for their in-progress graphic designs. Our findings indicated that while most designers approached the community for feedback only once when their design was nearly finished, their likelihood of returning for feedback on subsequent iterations was positively correlated with the feedback’s length and number of thought-provoking critiques they received, and was negatively correlated with the amount of praise they received. The community appealed to designers primarily due to its organization of feedback into reply threads and the potential to solicit feedback from a large, diverse audience. This work contributes a deeper empirical understanding of how characteristics of feedback and its presentation relate to iteration, as well as findings that

can inform the design of online feedback exchange platforms to support iteration. A paper reporting the results of this study was published in *ACM Creativity and Cognition*, 2017.

Together, the contributions discussed above map out how the interfaces present throughout the feedback exchange process link to the way the feedback is composed, explored, interpreted, and utilized. This map provides a strong foundation for helping designers of feedback support tools and interventions leverage interface design to facilitate communication between creators and their feedback providers. My dissertation contributes to improving the quality of feedback exchange for iterative processes for open-ended work.

Chapter 2: Related Work

I situate the contributions of this dissertation in the existing bodies of research investigating the feedback needs of creators and developing techniques for improving the feedback exchange process. In this chapter, I summarize literature surrounding feedback interpretation and composition roughly organized by their pertinence throughout the remaining chapters. In the first section below, I discuss key challenges creators face when interpreting feedback, how structuring feedback or the interpretation thereof can help address these challenges, and how visualizations can help creators interpret unstructured feedback. In the second section, I summarize how and why creators engage in online feedback exchange, strategies for helping providers write feedback to support creators’ feedback exchange needs, and techniques for facilitating feedback exchange at scale.

2.1 HELPING CREATORS INTERPRET FEEDBACK

2.1.1 Challenges of Interpreting Feedback

Creators may not find feedback helpful if they are unable to interpret it, even if it is of high quality [9, 16, 32]. This is particularly true of novice creators, who are less likely than experts to question and reconcile inconsistencies in the feedback they receive [33]. Interpretation challenges may stem from a creator’s own cognitive barriers and behaviors. Winstone et al. [34] identify four main cognitive barriers to feedback recipience: 1) awareness of feedback’s meaning and purpose, 2) cognizance of strategies for interpreting feedback, 3) agency to implement those strategies, and 4) volition to engage with and implement the feedback. Several works reinforce these findings by demonstrating that gaps between a feedback provider and recipient in terms of domain knowledge [35, 36, 37], expertise [31, 38], and academic qualification [31, 39] may all inhibit interpretation. McCarthy [40] found that anonymity made creators less apprehensive and more receptive towards the feedback they received online. Cook et al. [41] demonstrated that although students who asked for specific, actionable suggestions for improvement made stronger revisions to their work than students who asked for other kinds of feedback, students rarely asked such guiding questions when soliciting feedback. Other studies have shown that a defensive mindset can also inhibit the reception and interpretation of feedback [17, 42, 43].

A creator’s ability to interpret feedback may also be influenced by attributes of the feedback itself. Wu and Bailey [21] found that presenting negative feedback at the end of a critique

rather than the beginning or middle improved creators' affective states, perceptions of the feedback and its provider, revision depth, and story quality. Butler [4, 44] and Lipnevich [19] showed that grades and praise inhibited students' improvement on creative works when presented alongside constructive comments. Winstone et al. [34] found that students had difficulty interpreting and implementing individual pieces of feedback when they were unsure how it related to their personal goals or when the feedback contained complex language. Gibbs and Simpson [45] found that students often skimmed large amounts of feedback rather than thoroughly reviewing it in its entirety. Finally, Foong et al. [33] showed that novices in particular may struggle to resolve feedback conflicting with their own frames of reference, arguing that online feedback exchange systems should provide additional structure and support for sensemaking.

The works above identify how factors such as showing grades with feedback, reorganizing feedback, and presenting conflicting feedback can influence the way creators interpret that feedback. However, these works do not identify how these factors link to feedback perceptions and utilization from an interface design perspective. My dissertation examines how interfaces that present a quality score (Chapter 4) and organize feedback at different levels of detail (Chapter 5) influence creators' perceived fairness, investment, and helpfulness of the constructive feedback they receive. I also relate these perceptions to the types and depth of revisions creators perform in response to the feedback, as well as the effort creators put into these revisions. Chapter 5 additionally weighs providers' perceived difficulty of using interfaces for composing feedback at different levels of detail against creators' perceived value of the resulting feedback.

2.1.2 Structuring Feedback Interpretation

Research for facilitating feedback interpretation typically focuses on cognitive strategies that structure creators' interpretation processes or tools that add structure to the feedback itself. One thread of research emphasizes self-directed cognitive interventions to help the recipient interpret feedback. For example, Jackson et al. [28] demonstrated that having students draft action plans for the feedback they received on their coursework increased their subsequent feedback utilization and grades. Yen et al. [16] found that asking designers to perform an explicit reflection activity after feedback review led to an increase in perceived quality of the revised designs compared to reviewing feedback without reflection. A subsequent study showed that asking designers to paraphrase feedback they received improved their comprehension of difficult words and concepts in the feedback, and led to more effective revisions than when no paraphrasing was performed [46]. Wu et al. [20] found that presenting

effort and expertise cues with a piece of feedback could significantly influence a creator’s perceived quality of the feedback. Wu later found that having users perform coping activities such as expressive writing or reflection after reviewing negative feedback increased their resilience to harsh criticism [17]. Finally, Cook et al. [47] found that reflection could help users better recall their design goals, question their choices, and prioritize revisions.

Another thread of research leverages interaction design to help users navigate large collections of structured feedback. For example, the Voyant [7] tool generates structured feedback for graphic designs from a non-expert crowd and presents the feedback using word clouds, interactive graphs, and annotations to help users extract high-level themes from the feedback (Figure 2.1). A classroom study found the crowd feedback presented in Voyant helped students improve the quality of their graphic designs, and was perceived as more interpretative, diverse, and critical compared to free-form feedback generated for the same designs [15]. CrowdCrit [9] structures and aggregates crowdsourced design critiques through an interactive visualization showing the self-rated expertise of the providers as well as quality ratings for the graphic design along several design principles. The authors found that users reported noticing more issues and producing better designs using aggregated crowd critiques than when using generic feedback.

My dissertation intersects research on cognitive interventions and interaction design by comparing how different interfaces for presenting feedback impact the way creators perceive and interpret that feedback. Chapter 5 compares how interfaces that show feedback at different levels of detail impact creators’ perceptions of both their feedback providers and the feedback itself. I link these perceptions back to how providers’ perceive their composition process at each level of detail, and build an emergent framework for selecting composition and presentation interfaces based on creators’ goals. Chapter 6 examines how visualizing metadata about feedback’s topic, opinion, and provider influences creators’ goals and strategies when interpreting large collections of feedback in practice. I investigate the unique interpretation processes that creators report employing when interacting with the visualization, and synthesize insights from these reports to inform the future use of structured visualization to support interpretation.

2.1.3 Interpreting Unstructured Feedback

The tools described in the previous section aid feedback comprehension by structuring either creators’ processes for interpreting the feedback or the generation of the feedback itself. However, structuring how feedback is processed and generated is not always feasible nor desirable. Several works have explored using visualization techniques to help creators and

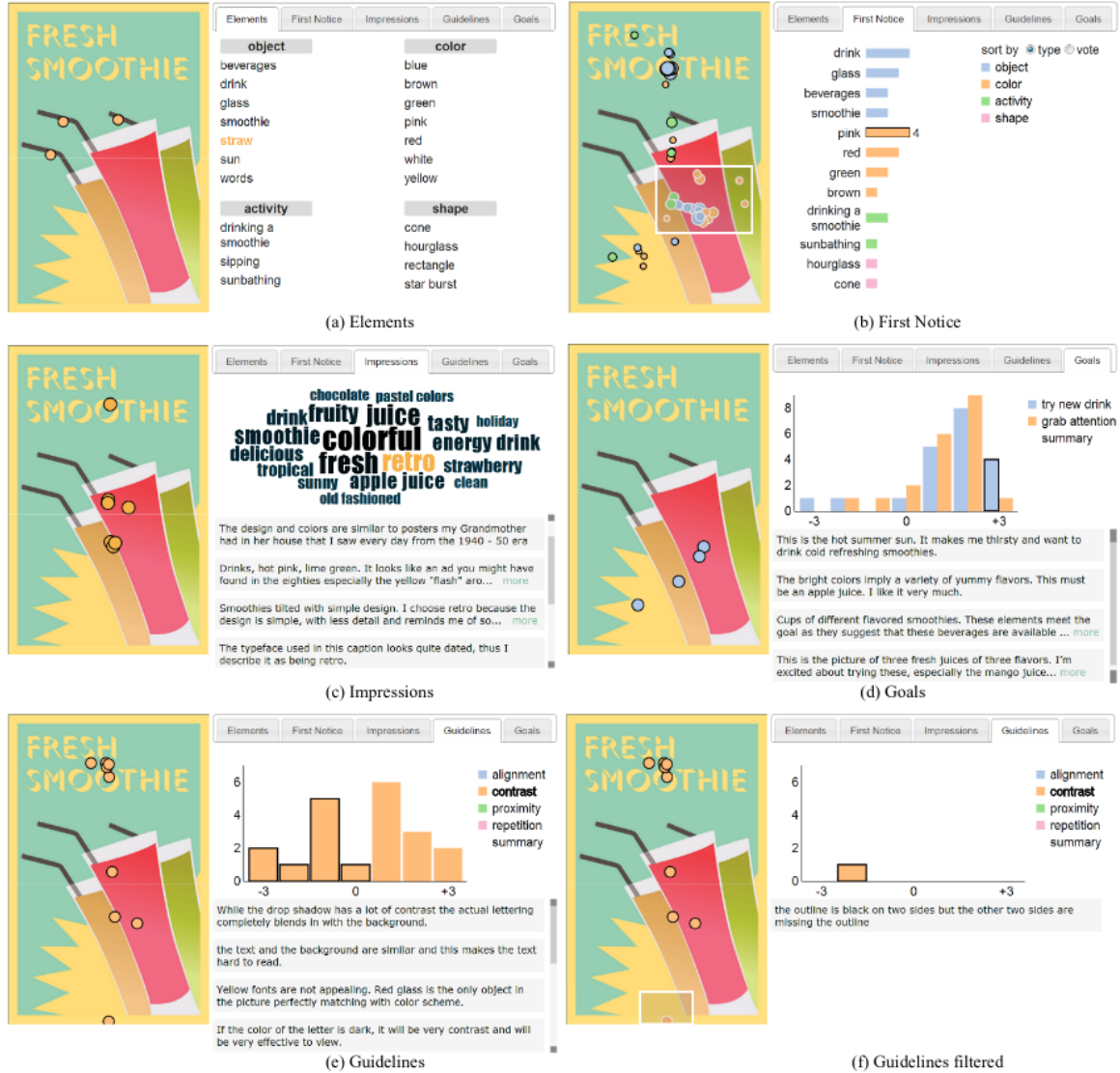


Figure 2.1: The Voyant system [7] provides multiple interactive visualizations summarizing crowdsourced graphic design feedback based on individual design elements and principles, the creator’s design goals, and users’ first impressions of the design. Visualization tools such as Voyant are useful for helping creators interpret feedback that is already structured in some way, but do not help creators interpret unstructured feedback. Chapter 6 of this dissertation explores how an interactive visualization tool impacts the processes and techniques creators leverage to navigate and interpret large collections of unstructured feedback.

decision-makers interpret the unstructured feedback they often receive in practice.

CommunityPulse [48] helps civic leaders make sense of written feedback gathered from community members in response to civic design proposals. Among other tasks, the tool allows civic leaders to interactively explore the distribution of sentiment in comments written by the community members for each proposal. The authors found that the use of the tool

reduced the time and expertise required for community input analysis.

Review Spotlight [49] produces summaries of user-generated reviews using adjective-noun pairs, and allows users to explore the contexts of these pairs in greater detail within the reviews. Users of the tool were able to form detailed impressions about restaurants and decide between two options faster than they could when using traditional review web pages.

Crowdboard [50] projects real-time crowd feedback onto draggable sticky notes on a virtual whiteboard to help creators visualize and organize feedback on their ideas during brainstorming sessions. The authors found that creators valued the real-time crowd feedback the tool provided and incorporated it into their discussions, generating more creative ideas than creators who utilized a traditional whiteboard.

Finally, Unakite [51] collects, organizes, and displays alternative solutions to programming problems in terms of their relative tradeoffs. A user study showed that the tool reduced the cost of capturing tradeoff-related information by nearly half, and that developers understood these tradeoffs about three times faster.

Studies of the tools described above indicate that creators and decision-makers find tool support helpful for interpreting unstructured feedback. However, these studies do not explore how such tools influence the interpretation goals and processes creators develop for accomplishing tasks unique to interpreting unstructured feedback, such as prioritizing suggestions or identifying recurring criticisms. Chapter 6 of this dissertation explores what goals creators have when interpreting feedback, how they leverage an interactive visualization to pursue these goals, and how effectively different features of the visualization help them accomplish these goals.

2.2 HELPING FEEDBACK PROVIDERS MEET CREATORS' NEEDS

2.2.1 Expectations and Strategies for Online Feedback Exchange

Creators often approach online design communities and platforms for feedback on their in-progress creative works. Xu & Bailey investigated the expectations and motivations of users participating in an online critique community in digital photography [52]. They found that designers' participation in a community could enhance the perception of their work, provided they received feedback of sufficient quality and quantity. They also found that designers often approached online communities with the goal of obtaining quick feedback from members with comparable or greater design experience. Marlow & Dabbish [53] found designers derived professional benefit from sharing and promoting their work in an online graphic design community, and were able to improve their creative skills by reviewing and

mirroring the practices found in the design work showcased by others. They also found that designers, especially novices, valued the work-in-progress section of the site to receive quick feedback on their work. Other works have shown that crowdsourced feedback may approach the quality of expert feedback when taken in aggregate [9, 13] and can be especially helpful to projects targeting specialized audiences [54].

Considerable research has explored strategies for soliciting better feedback from online platforms. Yen et al. [55] demonstrated that soliciting feedback from socially, financially, and intrinsically motivated crowds each yielded feedback of comparable quality but different valence and content. Hui et al. [56] suggest that leveraging students' collective social networks could help more evenly distribute the volume of feedback received in design courses. Cheng et al. [10] found that creators were able to solicit higher quality feedback from crowds by signaling as a novice, critiquing their own designs, and providing variants on their designs. Prior work has also shown that recipients' perceptions of feedback quality may increase when prompted to ask directed questions about their work to feedback providers [7, 8, 41].

The works above identify why and how creators engage in online feedback exchange, as well as factors that creators consider when seeking feedback (such as the motivations and expertise of their feedback providers). My dissertation extends these studies by exploring how additional factors influence feedback exchange from the perspectives of both creators and their feedback providers. Chapter 5 investigates how providers' perceptions of composing feedback at different levels of detail compare with how creators interpret feedback at each level. Chapter 6 explores how interactive visualization techniques facilitate interpretation, and how seeing the feedback visualized influences a provider's subsequent feedback composition. These studies expand upon prior work by contrasting the goals and expectations of creators with those of their feedback providers, and highlight opportunities to support the needs of all parties involved in the feedback exchange process.

2.2.2 Composing Feedback for Individual Creators

Producing effective, high quality feedback is the primary goal of feedback composition. Kluger and DeNisi's prominent Feedback Intervention Theory [57] proposes that effective feedback contains cues directing attention away from meta-task processes and towards task-motivation processes, task-learning processes, and goal-setting interventions, all while minimizing cognitive load. Similarly, Sadler [58] suggests effective feedback helps recipients develop strategies to modify their work and improve its quality, while Lefroy et al. [59] concludes that effective feedback is specific, actionable, justified, and task-directed.

To satisfy the above criteria, one thread of research has explored various instrumentations

for helping providers write higher quality feedback. Toxtli et al. [60] found that displaying prompts reminding employers and customers not to criticize unfair factors outside a gig worker’s control resulted in fairer reviews. Cook et al. [41] demonstrated that guiding feedback providers with questions about what a creator could improve resulted in feedback that was more specific, actionable, and critical. Hicks et al. [12] found adding a numeric scale to open comments elicited reviews with more explanations but of lower quality, while decomposing reviews into shorter stages elicited more diverse feedback.

A different thread of research aims to help providers write effective feedback by bridging gaps between creators in terms of their expertise or domain knowledge. Templates [61, 62, 63] and rubrics [13, 23, 24] have been shown to help non-experts write feedback comparable in quality to that of domain experts. Shannon et al. [64] showed that live rubric-based peer review during in-class presentations helped students produce immediate, relevant, and diverse feedback, with over 80% of comments being rated as helpful by their peers. Another class of techniques leverages comparisons with exemplary works to bridge gaps in expertise. Kang et al. [6] (see Figure 2.2) showed that feedback providers wrote more specific, actionable, and novel feedback when asked to select visual examples of designs relevant to their feedback. Cambre et al. [11] demonstrated that students who composed feedback for one design while contrasting it with a second design wrote feedback that received higher expert ratings than feedback from students who did not review contrasting designs.

The works above test the influence of several interventions on the perceived quality of feedback, but do not examine how recipients actually use the feedback. Chapter 5) of my dissertation explores the effectiveness of four interfaces for composing feedback at different levels of detail from the perspectives of both the feedback provider and feedback recipient. From the provider’s perspective, my work explores the time and effort required to compose feedback at each level of detail, the perceived limitations of composing feedback at each level, and how helpful a provider ultimately perceives their feedback will be to the recipient. From the recipients’s perspective, my work explores the creator’s own perceptions of the feedback’s helpfulness, fairness, and personalization, as well as how these perceptions correspond to the depth and quality of revisions a creator performs on their work.

2.2.3 Composing Feedback at Scale

Techniques for scaling feedback composition become increasingly necessary as more creators rely on online platforms for feedback. One prominent technique involves reusing or reappropriating feedback for use by others beyond the original intended recipient(s). In programming education, Moghadam et al. [65] developed a methodology for providing pro-

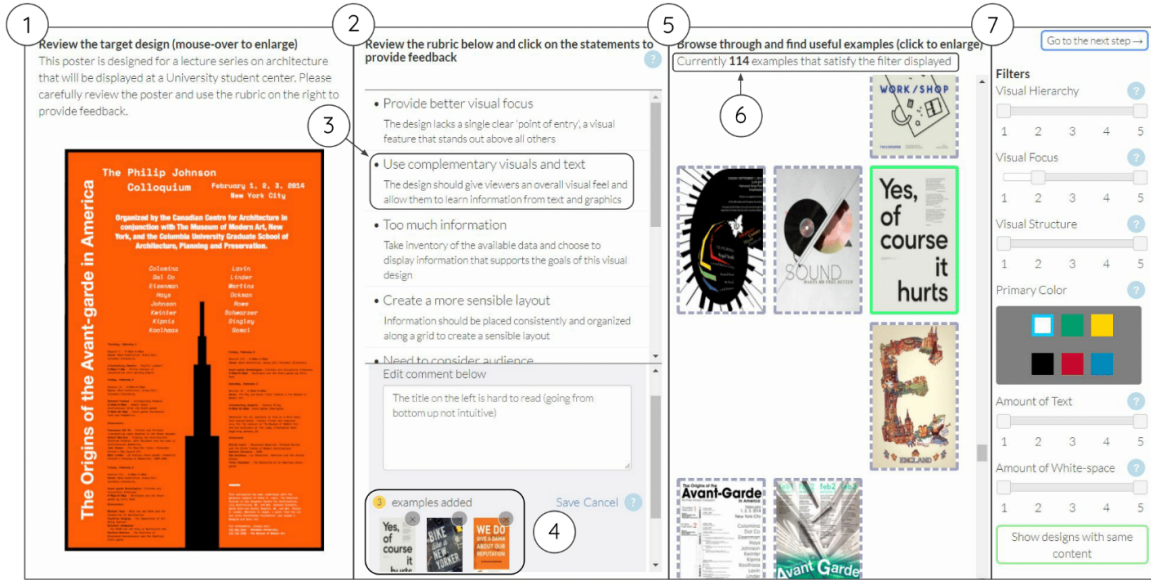


Figure 2.2: Paragon [6] assists feedback providers in composing high quality feedback by leveraging comparisons to exemplary designs. The composition interface shows the feedback provider the target design and explanation (1), then has them select feedback from a rubric (2), edit the feedback text as they see fit (3), and select design examples relevant to their feedback (4-6) using several filters (7). The authors of Paragon investigate how metadata can affect feedback composition, but do not link these effects back to how creators perceive the feedback or subsequently revise their work. By contrast, the methodologies of the studies presented throughout this dissertation examine and link several stages of feedback exchange (see Table 1.1) ranging from a provider planning their critique to a creator incorporating the feedback into their work.

gramming students with automated feedback on coding style based on prior exemplary submissions. Glassman et al. [66] built a user interface for generating variable-name feedback for code, while Head et al. [67] prototyped a system for repurposing feedback among students whose code contained similar bugs. For graphic design, Ngoon et al. [68] prototyped a system for curating reusable snippets of feedback on design projects, while Cambre et al. [11] leveraged structured comparisons to solicit feedback for multiple designs simultaneously.

Another scalable option is to use automated assessments, which allow a small number of providers to give feedback to more recipients than otherwise possible using traditional personalized feedback. Existing systems for automatically assessing open-ended work typically 1) use heuristics to assign quality scores to individual components of the work, and 2) select constructive feedback corresponding to these scores from a pool of pre-authored written comments. Mahajan uses this pattern in StudyCrafter to produce critiques of interactive narrative projects [69], while Bharadwaj incorporates this pattern into Critter’s AutoQA tool

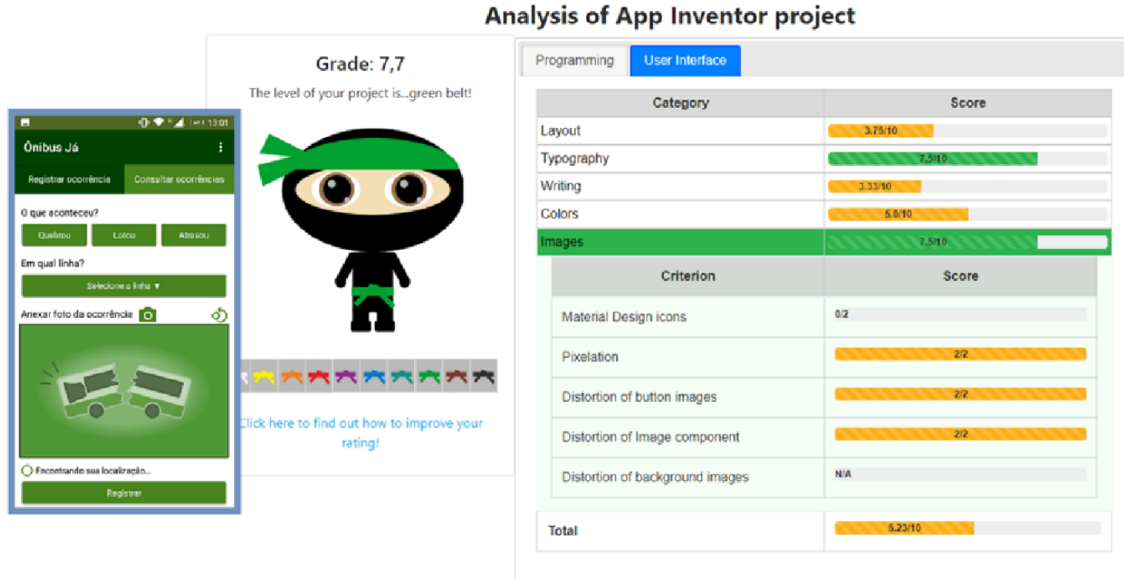


Figure 2.3: The Codemaster [72] system automatically generates feedback on the visual design of Android apps using heuristics based on common user interface design principles. This feedback is summarized through a visualization and used to assign an overall quality score to a project. Codemaster is representative of a class of scalable automated systems that provide scores and pre-authored comments on creative works. However, prior work has reported mixed findings regarding whether scoring creative works is beneficial or detrimental to helping creators interpret constructive feedback. Chapter 4 of my dissertation explores how quality scores and pre-authored constructive comments interact to influence creators' perceptions of the feedback they receive and subsequent revisions to their creative works.

for assessing the quality of web designs [70]. In minor variations of this pattern, Cutumisu uses binary quality scores to generate automated feedback on poster designs with Posterlet [71], while Solecki uses rubric-based feedback in the CodeMaster system for evaluating Android apps [72] (Figure 2.3).

In contrast to the domain-specific solutions developed in the preceding works, a separate thread of research has explored more general strategies for automating assessments. Foltz [73] and Pardo [74] leveraged learner analytics to offer personalized recommendations for students based on their interactions with digital systems. Malik [75] developed a methodology for generating substantiable grades on students' work using an inference network trained on examples synthesized from real solutions. Kumar [76] evaluated the potential of using deep learning to help produce rich grading algorithms in automated essay scoring systems.

My dissertation extends research on scalable composition by examining how different interfaces influence feedback composition costs and techniques for open-ended creative projects. Chapter 4 investigates how quality scores and pre-authored written comments

typical of automated assessment systems interact to influence creators' perceptions of feedback and their subsequent revisions. Chapter 5 examines how the time, effort, and skill sets required to compose feedback at different levels of detail relates to the quality of the feedback as perceived by both the feedback providers and recipient.

Chapter 3: Preliminary Work: Exploring How Designers Approach Iteration in an Online Critique Community

Feedback is essential to any iterative, creative process, helping creators identify flaws in their designs and gaps between their intentions and others' perceptions of their work. While creators have turned increasingly to online platforms for feedback, little research has explored when and how creators engage with other users of these platforms to solicit feedback as part of their iterative process. Even less research has investigated if and how the feedback that creators receive from these online sources links back to iteration and revisions to their creative works. To address this gap in the literature, I conducted a field study of how novices approached sharing and soliciting feedback for their in-progress work within an online design community.

In this study, I explored how novices approach iterative design practice in three active forums targeting novice design critique. I surveyed users ($n=38$) from the community and analyzed a large publicly available corpus of projects ($n=3,730$) and comments ($n=29,412$) from the platform to determine 1) why designers chose this platform instead of (or in addition to) other platforms, 2) when and how often they sought feedback on their work using the platform, and 3) how the feedback they received ultimately affected their creative process and outcomes. The study described in this section highlights the importance of how the design of user interfaces present throughout the feedback exchange process may encourage or deter creators from revising their work. The results from this study further motivated me to explore how user interfaces influence creators' processes for interpreting, navigating, and operationalizing feedback.

3.1 INTRODUCTION

Iteration is essential for producing creative solutions and gaining confidence in one's creative ability [77, 78]. Through iteration, designers learn to perform a series of content revisions prompted by feedback from an external audience [79]. The feedback enables the designer to see problems with their proposed solution, learn about the design problem, and gain insight for improving the work [80]. Although seasoned designers often have the skills and the resources necessary to iterate effectively, many designers do not, and face the challenge of receiving timely, helpful feedback [8].

For designers with limited resources, particularly novices, online design communities can serve as a source of feedback and creative inspiration for their projects [52]. Participation generally does not require financial resources or social capital, and often requires only creating

an account. These communities allow designers to connect online with diverse feedback providers who are motivated by a mutual interest in design or the topic of the projects. Even experienced designers can benefit from online communities, as they enable them to showcase and promote their work, learn about modern design trends, and refine their creative skills [53].

Although researchers have studied feedback generation in online design communities [52, 53], there are still important open questions regarding how the feedback requests are integrated into iterative design practice. For instance, when and how often do designers iterate on their projects using the feedback received from an online community, what characteristics of the feedback impact iteration, and to what degree do the projects improve?

We report results from a mixed-methods study of how designers approach iteration in three graphic design critique forums in Reddit. We chose Reddit because it contains some of the largest and most active public forums for design critique. We collected a large corpus of design projects (3,730) and critique comments (29,412), and applied heuristics to determine when designers posted iterations for their projects. These heuristics were developed by observing common practices for how iterations were represented within the community. We measured how often designers iterated on their posted projects, and statistically modeled which characteristics of the feedback received (e.g., the length and number of responses, the valence of each response, and the categories of critique discourse referenced in the responses) correlated with subsequent design iteration. The quantitative measures were complemented by a structured (N=21) and an open-ended (N=17) survey posted to the community. The surveys inquired about participants' motivations for and experiences with receiving feedback from the community into their design process, and asked participants to quantify and describe their iterative process.

Our study contributes three findings. First, we found that more and longer comments, comments with fewer positive statements, and comments that contained more thought-provoking statements were all predictive of iteration. This indicates that the type of feedback received online affects how often designers share iterations of their work. Second, we found that only a single design was posted for feedback for 79% of the projects in our data set, while two or more iterations were posted for the remaining 21%. The survey results indicated that designers were approaching the community near the end of their process, leaving less time and possibly less desire to continue iterating on their work. Finally, we found that designers posted their projects in the community with the expectation that they would receive quick, high quality feedback from a large, diverse audience. This extends findings from a prior study of professional development in an invitation-only design community [53] to an open community that targets novice design critique.

Our work makes two contributions to the HCI community. First, our results contribute deeper empirical understanding of the characteristics of feedback received online that relate to design iteration and how designers incorporate the use of online communities into their creative projects. Second, our findings can inform how online communities and crowd-based design services (e.g., [7, 8]) can better represent projects and generate the types of feedback that promote iteration. We believe these contributions will enable online communities to serve as more effective resources for creative design projects.

3.2 RESEARCH QUESTIONS

This study was designed to answer the following questions:

- **RQ1:** How many iterations on a design project does a designer typically post to an online community for feedback?
- **RQ2:** When a designer posts a revised design, how deep are the changes, and does the quality of the design improve?
- **RQ3:** What characteristics of the feedback (e.g. valence, category of discourse, or total comments) correlate with the decision to post subsequent design iterations?
- **RQ4:** When in the design process does a designer typically share a design with an online community?
- **RQ5:** What are the perceived strengths and weaknesses of iterating with the feedback received from an online design community (compared to other sources)?

Though not exhaustive, these questions were posed to learn about how designers approach online communities for feedback, to measure potential benefits of online iteration, and to gain insight into how online communities could further improve feedback exchange and iterative design practices.

3.3 METHODOLOGY

A mixed-methods study was conducted to answer our research questions. The first three questions were answered by analyzing a data set collected from an online community that targets novice designers (but open to designers at all skill levels). The community was therefore especially interesting because novices may not be as aware of the need to iterate as

more experienced designers. The final two questions were answered by posting two surveys to the community.

3.3.1 Online Community Studied

We collected designs and the associated feedback from Reddit (<http://reddit.com>). Launched in 2005, Reddit is a publicly accessible platform for discussing items of interest. The site originally targeted news, but has grown to include a variety of topics, including design. Topics are organized into sub-Reddits, and each sub-Reddit lists discussion threads ordered by a combination of popularity and creation time.

We chose Reddit because it has an active, large, and diverse user base engaged in design feedback exchange and, unlike other design communities [52, 53], participation does not require invitations, portfolios, status, or payment. Most designs posted to the sub-Reddits studied are visual designs, including graphic, web, and interaction designs. Examples include personal websites and logos, portfolios, T-shirt designs, business cards (self-employed), and non-commercial apps. From our own inspection, these projects generally arise from designers pursuing their own interests, learning goals, or job responsibilities. From the descriptions in the initial posts, the focus was typically on producing design solutions rather than solely learning about the design process or strategy; e.g., typical project posts include:

“Any help is appreciated! Just starting up so we don’t have that many products yet. What do you think of the logo? Would it be better bigger and without the palm tree?”

“Hi guys, here is my draft for a poster advertising a show I’ll be playing later this year. I would like the poster to be a combination of professional and trendy, and appeal to a diverse audience.”

Although the sub-Reddits we studied were open to designers of all skill levels, their target audience was novice designers. For example, the most active sub-Reddit we studied, /r/design_critiques, states its purpose is to “Help new and amateur designers improve their designs.”

On Reddit, a designer initiates discussion around her design by creating a thread in the desired sub-Reddit. The initial post typically contains a title, and description of the design and goals of the project. It may also include links to external images or to a live Web site. Community members can then comment on the design and reply to each other through a

discussion interface. Fresh designs typically receive fast attention but fade as the display algorithm places new posts at the top of the thread list, as also observed in [81, 82, 83].

To post a revised design, the designer can create a new thread with a link to the revised design (and the prior thread if desired), edit the original post with a link pointing to the revised design, or post a new comment in the original design’s thread with a link to the revised design. The last option is most common, as it is more noticeable than an edit and, unlike a new thread, it preserves the feedback history.

3.3.2 Design + Feedback Data Collection (RQ 1-3)

We developed a script using the Python Reddit API Wrapper (PRAW). The script crawled three popular sub-Reddits for design critique: /r/design_critiques, /r/Logo-Critique, and /r/logodesign. These sub-Reddits were chosen due to their similar purpose for visual design critique, target audiences, and norms. The script collected data for two one-month periods (separated by six months). The script downloaded the 1000 most recent threads in each of these sub-Reddits, including the design images, comments, timestamps, and user ids. This was the maximum data allowed by the API.

For the initial post and any subsequent comments by the designer in the thread, the script parsed the content for Web links. If found, the script downloaded the linked images or, if it pointed to a live Web site, rendered the linked page and captured it. These images were the designs in our data set. It was rare for designers to request feedback without linking to a design image, or to link to content that were not designs.

A challenge was detecting when designers posted revisions of their projects. By observing common practices on the site, we developed three heuristics: 1) the designer who created a thread replied with a comment containing an image link; 2) the designer edited the link in her original post; or 3) the designer created a thread that contained a link to an image and a link to a prior thread created by the same designer. Links to live sites were also challenging because the designer could update the site without editing the link. To detect this case, our script compared the rendered image of the page to the last capture for the link, if it existed. The comparison was performed using a well-known algorithm (MD5 checksum). If different, the image was categorized as a design iteration. Iterations were only detected within sub-Reddits because we observed that iterations were rarely split between them.

The timestamp of a comment was used to associate it with a corresponding design iteration, which was the last design iteration detected prior to that timestamp. We inspected samples of the data to confirm the design iterations and that comments were being associated with the correct iteration.

Based on our observations, if a designer posted more than one iteration, it unfolded within a short window (e.g. a few days). A month of data collection should therefore capture most of the iterations on a project shared to Reddit. However, it misses iterations that only partially overlapped the edges of data collection, as well as any iterations not shared within the community; we revisit the latter point in the Discussion.

3.4 RESULTS

Tables 3.1 and 3.2 summarize our data set. The data contained a total of 3,730 projects from 2,866 designers, and 29,412 comments on those projects (7,855 comments came from the designers, not including the initial posts). A design received about six comments on average, which is more than other studies of design critique communities have reported [52]. The average number of comments per design iteration was largest for /r/logodesign ($\mu=9.2$, $\sigma=14.9$), and least for /r/design_critiques ($\mu=3.9$, $\sigma=3.5$), with /r/Logo_Critique in-between ($\mu=4.7$, $\sigma=3.7$). The complete set of comments received by a design were, on average, received within 24 hours of the initial post. The summary data shows that the designer and community often engage in back-and-forth discussion, reminiscent of face-to-face critique [84].

Sub-reddit	Projects	No it.	2 it.	3 it.	4 it.	5+ it.
/r/design_critiques	1485	1238	196	42	6	3
/r/logodesign	1173	893	226	36	11	7
/r/Logo_Critique	1072	832	186	36	13	5
Total	3730	2963	608	114	30	15

Table 3.1: The number of projects collected for each sub-Reddit in the data set. The subsequent columns show the number of iterations shared for each. As shown in the No Iteration (it.) column, only one design was posted for 79% of the projects.

Sub-reddit	Total Comments	By the Designer	By the Members	Avg. (SD) Per Project
/r/design_critiques	8502	2769	5733	3.9 (3.5)
/r/logodesign	13787	2987	10800	9.2 (14.1)
/r/Logo_Critique	7123	2099	5024	4.7 (3.7)
Total	29412	7855	21557	5.9 (7.1)

Table 3.2: The total number of comments collected from the sub- Reddits, by the designers and community members. Despite high variance, nearly all designs received feedback.

3.4.1 Iteration Analysis (RQ1)

Table 3.1 shows the distribution of how many projects share different numbers of design iterations for feedback. For instance, the No Iteration column shows the number of projects for which the designer posted only a single design for feedback and did not return, the 2 Iterations column reports the number of projects for which the initial design and one revision were posted, and so on. We ran a script to count the total number of projects in each group to determine how often designers iterate. Despite the widely-evangelized benefits of iteration [15, 79, 85, 86, 87], of the 3730 projects analyzed, only about 1 in 5 (20.6%) posted more than one iteration, and only about 1 in 23 (4.3%) of the projects received three or more iterations; these rates were similar across all three individual sub-Reddits.

We also analyzed how designers approached iteration based on the three heuristics identified in the Method section. We found that designers performed 863 (85%) iterations by replying to their original threads, 139 (14%) by editing their original threads, and only 15 (1%) by creating an entirely new thread for their revised designs.

Some possible explanations for the observed rate of iteration are that the designers who approach this community for feedback are unaware of the benefits of iteration, that they are approaching it too late in their design process, that the platform does not encourage iteration, or that the content of the design feedback received may not prompt iteration. Our later analysis will examine which, if any, characteristics of the feedback correlate with posting two or more iterations. We also return to this issue in the Discussion.

3.4.2 Perceived Quality and Iteration (RQ2)

To determine if designs improved when iterating based on community feedback, we separated the projects with only one iteration from those with two or more iterations. We then randomly sampled 300 of the projects with two or more iterations, and filtered out the ones lacking either a description of the designer’s goals or community feedback, leaving 102 designs. We felt this data set was sufficient to produce representative results. For each project, we selected a design image from the initial shared design and the final shared revision. If a designer posted multiple images for the first design or final iteration (a rare occurrence), we selected the one that we felt best represented the designs at that stage.

For each design, we recruited participants to judge the quality of the designs relative to the designer’s stated goals. The judging consisted of three categories of rating tasks:

- Rate how well a single design satisfied the designer’s goals on a 7-point scale (1=does not satisfy, 7=satisfies the goals). The goals were extracted from the designers’ original

Slight revision (rating=2)	Significant revision (rating=6)
 <p>Initial design for CEVSOC pizza night poster. It features a wooden background with a pizza slice graphic in the center. The text 'CEVSOC presents' is at the top, followed by 'END OF SEMESTER PIZZA & BEER NIGHT'. Below the graphic, it says '6 P.M. THURSDAY 29 OCTOBER', 'LEVEL 5 DESIGN STUDIO', and 'CIVIL & ENVIRONMENTAL BUILDING'. There are also images of basil, tomatoes, and a bowl of cheese.</p>	 <p>Initial design for That One Geek Productions logo. It features the text 'That One Geek' above a circular logo containing a stylized house with glasses, and 'Productions' below it.</p>
 <p>Final design for CEVSOC pizza night poster. It is identical to the initial design, featuring a wooden background, a pizza slice graphic, and the same text and images.</p>	 <p>Final design for That One Geek Productions logo. It features a circular logo with a stylized house and glasses, followed by the text 'That One Geek Productions'.</p>

Table 3.3: Examples of the first and final iterations for two projects with slight (left column) and significant (right column) perceived revision between the iterations.

post and displayed at the top of the task screen. The design image was shown below. The initial design and final iteration were rated as separate tasks.

- Rate the degree of difference between the two design images on a 7-point scale (1=the same, 7=completely different) and list the key differences between them (see Table 3.3). The designs were shown side-by-side on a single task screen, with the placement

randomized.

- Select which of the two design iterations better satisfied the designer’s goals. The two design images were placed side-by-side, with random placement. The goals were shown at the top of the task screen with the images below.

We piloted and revised the task screens to give simple, clear instructions to maximize response quality. We recruited participants from Mechanical Turk. Three participants were recruited per instance of the first task category, and five were recruited per instance of the other two categories. The task configurations limited recruitment to the U.S. to reduce language barriers, and required 95% prior approval ratings to recruit workers with a history of quality work. Participants were remunerated \$0.60 per task for the first category and \$0.10 per task for the other two categories. We also recruited an independent rater to perform the AB-comparison task using an interface similar to that used on Mechanical Turk. The rater had several years of professional employment and education in industrial design, and ostensibly had much more design experience than the Turk participants.

From the results of the first task, we used the mean of the ratings for each design to indicate its perceived quality. Surprisingly, the means were the same for both the initial design ($\mu=4.773$, $\sigma=0.863$) and the final iteration ($\mu=4.775$, $\sigma=0.981$). A paired samples t-test confirmed that there was no difference between the two.

From the second task, the mean difference for the pairs of designs was ($\mu=3.96$, $\sigma=1.62$). A Kruskal-Wallis test showed no correlation between the degree of change between iterations and the change in perceived quality. Combined with the prior result, this analysis shows that designers were making revisions, yet the perceived quality did not improve.

For the third (AB comparison) task, we tallied the number of times each iteration was selected as better satisfying the designers’ stated goals. The votes were nearly equally split between the initial designs ($n=247$) and final iterations ($n=263$). A chi square test confirmed there was no preference between the iterations ($\chi^2(1)=0.50$, $p=0.479$). The lack of increase in perceived quality between the initial design and final iteration might be due to the quality of the feedback being insufficient to make substantial improvements, or to the designers’ inability to interpret and apply the feedback.

We analyzed the ratings from the independent rater for our AB task, and found that the final iteration ($n=68$) was selected over twice as often as the initial iteration ($n=30$), with 4 selected as being the same ($\chi^2(1)=14.74$, $p<0.001$). The fact that the independent rater detected a difference but the Turk participants did not might be due to the need for expertise to discriminate differences in quality [88].

As an additional check to ensure that the ratings from the MTurk participants were reliable, we eliminated responses from raters who spent fewer than 25 seconds (<25% of the average time) on the tasks. Analysis of the ratings from the remaining raters also failed to detect statistical differences.

3.4.3 Predictors of Iteration (RQ3)

We examined how the number and length of responses, idea units, valence, and content categories of the feedback correlated with the decision to iterate on the design. From our original data set, we randomly sampled 100 projects containing only one iteration and 100 projects containing two or more iterations, filtering out all projects that did not have comments (excluding comments from the project creators themselves). This left 86 projects with only one iteration and 97 projects with two or more iterations for analysis. We had a single coder partition each comment from the projects into individual idea units. An “idea unit” is defined as a coherent unit of thought consisting of a phrase, sentence, or group of sentences. Given the scale of the data set, we divided the idea units ($n=2,416$) among a team of four coders (including a member of the research team). All coders had experience with design critique and similar labeling tasks.

We had our coders categorize the valence of each idea unit. Valences were either “positive” (positive comments encouraging or praising the design or designer), “critical” (destructive criticism towards the design or designer), “neutral” (comments that were neither positive nor critical, including most constructive criticism), or “indeterminate.” Coders also categorized each idea unit according to an established taxonomy of critique discourse [89]. Categories included judgment, direct recommendation, brainstorming, process-oriented, identity-invoking, free-association, comparison, interpretation, and investigation. We also added the category “support” for idea units praising or encouraging the creator’s ongoing effort (see Table 3.4).

To test inter-rater reliability, a member of the research team labeled the valence and category of discourse of 100 idea units randomly sampled from each of the other three coders (300 total). There was a raw 77% agreement for the category of discourse (Cohen’s Kappa = 0.71) and 78% agreement for the valence (Cohen’s Kappa = 0.68), which are considered satisfactory for moving forward with analyzing the results [90, 91]. The categorized idea units were aggregated by project and iteration. We then counted the number of idea units falling under each category of discourse and valence, the total number of comments, the total number of idea units, and the total word count for each project revision. We used this data set for the regression models described below.

To avoid overfitting our data, we used three separate logistic regression models. In

Category	Description	Example Idea Unit	1 it.	2+ it.
Brainstorming	Feedback asking (often rhetorical) questions or making statements about imagined possibilities for the design.	Would you consider the neck of it a bit — even if it’d be inaccurate — to make it take up a bit less space?	5.3%	4.1%
Comparison	Feedback contrasting the design or design process with something else as comparison.	The second one is a little too internet- explorer reminiscent for me.	2.8%	4.5%
Direct recommend.	Feedback giving specific advice about a particular aspect of a design as a direct recommendation.	I would go all the way back to square 1 and work with letterforms.	25.7%	21.7%
Free association	Feedback that makes reactive, associative statements about the design as free associations.	I also like the fact that it looks like a lightbulb, as that is a pretty universal symbol of “eureka!”	3.2%	4.5%
Identity-invoking	Feedback pushing designers to consider themselves within the larger context of the design profession.	So, yes, it does show that you’re a beginner. That just means you have room to improve!	1.8%	2.5%
Interpretation	Feedback where someone reacts to what they saw and tries to make sense of the concept or product.	The first thing I see at a glance is mountains. Lots of mountains. Could be ski shop.	7.1%	6.5%
Investigation	Feedback that requests information about the design or the design process as investigation.	Have you researched other ramen restaurants logos?	4.8%	7.5%
Judgment	Feedback that is evaluative in tone and which often includes some form of interpretation while also conveying an assessment of the design.	The lines are very thin as well, but that is irrelevant as the design is just not good. Thickening the lines will not help.	40.4%	34.4%
Process-oriented	Feedback providing designers with insight or observations about the process that they might have used or could use to create the design.	Since it’s a restaurant I would strongly consider how the logo would look as a sign and how it will work with the interior/mood.	2.4%	7.2%
Support	Feedback expressing the provider’s support for the design creator.	Good luck either way!	3.2%	2.7%

Table 3.4: The taxonomy of critique discourse used to categorize the idea units. The two rightmost columns show the relative proportion of instances of each category for projects with only a single design (1 it.) and those with two or more iterations. Columns do not sum to 100% because idea units that did not fit into any of the above categories were omitted.

our first model, we used the idea unit content categories as the predictor variables, and iteration as the response variable. The model identified three categories as being predictors of iteration: process-oriented (coef.=0.750, $p < 0.001$), comparison (coef.=0.279, $p = 0.124$), and investigation (coef.=0.204, $p = 0.167$). Using chi-squared tests to compare the distribution of content categories among idea units for design projects with a single iteration to those with multiple iterations, we found similar results: projects with multiple iterations had a higher proportion of process-oriented ($\chi^2(1) = 25.97$, $p < 0.001$), investigation ($\chi^2(1) = 6.66$, $p = 0.010$), and comparison ($\chi^2(1) = 3.93$, $p = 0.048$) idea units. Conversely, projects with only one iteration had a higher proportion of both judgment ($\chi^2(1) = 8.85$, $p = 0.003$) and direct recommendation ($\chi^2(1) = 4.82$, $p = 0.029$) idea units. We did not find any significant difference in the proportion of free association, interpretation, brainstorming, identity-invoking, or support between the two groups.

Prior work found that designers with more expertise typically write more thought-provoking feedback (e.g., brainstorming and process-oriented) [89]. The higher amount of thought-provoking feedback on projects with multiple iterations might then indicate that those projects were receiving more attention from providers with more design expertise.

In our second model, we used valence as the predictor variable, and iteration as the response variable. The model identified both neutral (coef.=0.05, $p = 0.108$) and critical (coef.=0.09, $p = 0.148$) idea units as predictors of iteration, but not positive idea units. Chi-squared tests revealed similar results: overall, projects with multiple iterations had a significantly lower proportion of positive idea units than designs with only one iteration (19.3% vs. 24.9%; $\chi^2(1) = 10.77$; $p = 0.001$). Critical feedback reveals problems with the design and may prompt designers to iterate, whereas positive feedback may signal that the project is near completion and that further feedback is not necessary.

In our third model, we used the number of idea units, number of comments, and total word count per project as the predictor variables, and iteration as the response variable. The model identified word count as a predictor of iteration (coef.<0.01, $p = 0.002$). We applied a similar model on the entire corpus of design projects, filtered to only first iterations with at least one comment ($n = 3,406$); the analysis revealed that the number of comments was predictive of iteration (coef.=0.01, $p = 0.027$). One interpretation of these two analyses is that more comments and longer comments give designers more feedback to work with, exposes more potential improvements that can be made, and demonstrates a higher level of interest in and engagement with the work.

To determine if designers' engagement with the feedback providers affected iteration, we reran our models with designer engagement as a covariate. The ratio of the designer's comments to the total comments on a thread was calculated as a proxy for engagement. Since

designers typically reply to their own threads to represent iterations (see Iteration Analysis), not all designer comments were community engagement; consequently, this proxy measure is an upper bound for engagement. The models revealed that a higher ratio of designer comments to total comments was predictive of iteration (coef=5.29, $p < 0.001$ for the model related to category of discourse; coef=5.34, $p < 0.001$ for that of valence; coef=4.61, $p < 0.001$ for that of feedback quantity); the pattern of the results of the models were otherwise the same. This result suggests that increased engagement with feedback providers may also spur iteration.

Across all project iterations ($n=4,217$), we compared the number of comments received on the first and subsequent iterations. An ANOVA showed that initial iterations received more comments ($\mu=5.39$, $\sigma=8.65$) than subsequent iterations ($\mu=3.94$, $\sigma=4.22$; $F(1)=21.72$, $p < 0.001$). The fewer comments on subsequent iterations might be attributed to users being unaware the project has been updated (e.g., the site does not allow users to subscribe to posts). Another possibility is that community members may be choosing to direct their attention to other projects [92]. This re-affirms prior work showing that implementation choices, even subtle ones, can have a large influence on the distribution of feedback generated in an online community [52].

3.4.4 Perceptions of iterating online (RQ4, RQ5)

To complement our analysis of the forum data, we conducted two surveys. The first was open-ended, and asked designers about their motivations for and experiences with using Reddit for design feedback. It was posted to /r/design. The second was structured, and asked the designers about their iterative process. It was posted to /r/design_critiques and /r/logodesign, and was distributed to design-oriented mailing lists. We split the questions between two surveys to reduce the time required to complete each one, and used the responses from the first survey to help formulate the structured questions in the second survey. Both surveys required that a participant be at least 18 years of age and had posted a design for feedback to Reddit in the last six months. The surveys asked designers to describe a recent project posted to Reddit for feedback and to self-report their design expertise and demographics. We gave \$20 for participation.

In total, we received 38 responses to the surveys. There were 17 responses (four female, ages 18-45) for the open-ended survey and 21 responses (four female, ages 18-32) for the structured survey. For the structured survey, two responses (of 23, leaving 21) were eliminated due to not satisfying the survey criteria. There was no overlap in the participants. The majority of design projects posted by respondents of both surveys were logos and web site

designs; other projects included illustrations, portfolios, T-shirt designs, and flyers. The average self-rated expertise for the open-ended survey was ($\mu=3.3$, $SD=0.89$ on a 5-point scale), and the average for the structured survey was ($\mu=4.0$, $SD=1.4$ on a 7-point scale). We first report the data from the structured survey, and then draw from the open-ended survey responses to help explain the results from the structured survey and the prior analyses.

Structured Survey (n=21)

Table 3.5 summarizes the forced-choice questions and results from the structured survey. We note two interesting patterns in the data. First, the average reported number of iterations for a project (including those not posted to Reddit) was about six (Q1; $\mu=5.8$). However, the number of iterations posted to Reddit was less (Q2; $\mu=1.6$). For Q3, fourteen respondents reported that they posted their designs to Reddit near the end of their process and three reported they did so between the middle and the end. Two respondents reported sharing their designs at the beginning of the process, two at the midpoint, and one in-between. In sum, these results show that a large majority of the respondents posted one or two iterations to Reddit for feedback, typically near the end of their process.

Second, the respondents reported that iteration is important for their project (Q4; $\mu=6.0$), it is somewhat important to post iterations for feedback (Q5; $\mu=4.0$), the feedback received was reasonably good (Q7; $\mu=5.0$), and that Reddit is supportive of their process (Q8; $\mu=5.0$). Yet, consistent with the forum analysis, the respondents reported seldom posting more than two iterations to Reddit for feedback.

Open-ended Survey (n=17)

The open-ended survey asked respondents to describe their motivations for posting designs to Reddit for feedback, to identify other sites they considered posting their work and why they did or did not do so, to explain how the feedback they received helped (or failed to help) improve their work, and to explain why they would or would not use Reddit again as a source of feedback. We will use the notation Rn when referring to respondent n throughout this section. When asked about their motivations for posting their designs to Reddit for critique, there were three common responses (n=5 each). The first common response was that the potential of reaching a large audience was appealing:

“...getting comments from people on Reddit offer a wider range compared to my usual methods of asking friends and family what they think.” [R1]

Question	Mean (SD)
Q1. In total, how many design iterations did you create for the project?	5.6 (3.8)
Q2. In total, how many design iterations for the project did you post to Reddit for feedback?	1.6 (0.90)
Q3. At which stage in your process did you post the design to Reddit? (beginning, between beginning and midpoint, midpoint, between midpoint and end, near the end)	(see text)
Q4. How important was it for you to create multiple iterations for the project? (1=not important, 7=very)	6.0 (1.2)
Q5. How important was it to post multiple iterations to Reddit for feedback? (1=not important, 7=very)	4.0 (1.8)
Q6. How would rate the depth of changes made to your design based on the feedback received from Reddit? (1=minimal, 7=significant)	3.8 (1.8)
Q7. How would you rate the quality of the feedback received from Reddit? (1=low, 7=high)	5.0 (1.6)
Q8. How well does Reddit support your iterative design process? (1=not at all, 7=very well)	5.0 (1.4)

Table 3.5: The questions asked on the structured survey and the mean (SD) of the responses.

The second common response was that the site’s diverse user base brought fresh perspectives:

“...Reddit is a collective of individuals with varied taste and experiences of the real world, where the majority are not teachers. A teacher will critique on a different agenda, and a Redditor will critique on a personal bias, full disclosure, criteria. It’s more raw to get people outside your bubble, outside your country even, for design critique.” [R11]

The third common response was that the overall quality of the critique they received was high:

“Reddit is my go-to source for critique, advice, and general online discussion.” [R13]

When asked about other places they considered posting their designs, the respondents reported many different venues, but most stated that these venues were unsatisfactory for getting feedback. The most common was Facebook (n=5), followed by Behance, DeviantArt, and Dribbble (n=2 each); four respondents stated that they never considered posting their work to any other site. The respondents' reasons for ultimately not using these other sites were similarly diverse, and included the lack of a forum structure [R2,R5], lack of verbal critique [R3], lack of real world perspectives [R4], lack of anonymity [R11], and excessively harsh [R12] or impersonal [R7] critiques. This decision rationale indicates that Reddit has many unique qualities that make it attractive relative to social media and socially-oriented design communities as a platform for collecting design feedback.

When questioned about how the feedback helped them improve their designs, responses were more similar: nearly half of the respondents stated that users generally provided good critiques about the flaws of their designs (n=8), and over half stated that they provided good suggestions (n=9):

“Commenters pointed out a slight difference in sharpness/blurriness between the images and also suggested a tweak to the hue/saturation of one of the elements. Both are very slight, subtle tweaks but have a big impact on the end result.” [R15]

Not all respondents had good experiences, however. One respondent stated that Reddit was of little help [R12], and two stated that it was of no help at all [R1,R6]:

“I would either just have passive aggressive down votes or someone being very snarky with my work; not too many constructive things.” [R1]

When asked about why they would or would not use Reddit for receiving design critique in the future, respondents were divided. Those stating they would consider posting there again in the future cited the fresh perspectives of the users (n=7), the large user-base (n=3), the high quality critique (n=3), and the expertise of the users (n=2). Conversely, commonly cited reasons for not returning for feedback in the future included the low quality of the critique (n=3) and the lack of expertise among the users (n=3). One respondent stated outright that getting feedback was very “hit or miss”:

“They’re fairly good at feedback for the most part. Although it’s a hit or miss if you get 2 comments or you get 100. It depends on timing or likability or willingness of the other party.” [R11]

In sum, respondents found Reddit feedback mostly helpful for uncovering problems and getting suggestions for improvement, and felt that Reddit accommodates their iterative process more than other sites.

3.5 DISCUSSION

The goal of our study was to learn how designers naturally approach design iteration and feedback collection in online communities. Our quantitative results show that online design communities such as Reddit have immense potential for design critique: on average, designs receive about six comments, all within 24 hours, with each comment averaging 25 words and typically exceeding superficial statements such as “good job”. The values of these attributes exceed those reported in other studies of online communities [52], and other artifact-based discussion sites [93].

According to our survey respondents, the primary draw of using Reddit for design critique is the ability to reach a large and diverse audience who can offer fresh, authentic perspectives. The site’s users are ostensibly motivated by enjoyment of design or interest in a project, requiring neither social capital or compensation to provide feedback. Along with a low barrier to entry (e.g., invitations are not required), these factors make Reddit attractive for designers to post their projects. The majority of respondents to the surveys also indicated they were satisfied with the feedback received.

Our analysis showed that designers who posted revisions of their work received more process-oriented, comparison, and investigation categories of feedback. These types of feedback are typically more open-ended, questioning, and thought-provoking (e.g., “Perhaps you could try a couple different locations for the diamond”). This type of content may create more uncertainty about the design’s quality, thereby prompting recognition that further revision and feedback is needed. By contrast, designers who did not share multiple revisions received more recommendations for improvement and assessments of quality (judgments in the taxonomy). These categories may serve to validate the design (e.g. “Looks good”) or identify specific suggestions that can be directly accepted or rejected. Additionally, concrete feedback is characteristic of designers with less experience [89], potentially signifying that feedback from less experienced designers does not promote iteration. In either case, designers receiving this type of feedback may not see a compelling reason to share or even produce a revised design.

One way to solicit a desired balance of feedback categories is to allow designers to choose from defined rubrics that would scaffold the feedback generation process. Prior work has shown, for example, that rubrics can enable non-experts to give feedback comparable in

depth and quality to that of expert designers [13]. A designer could indicate her desire to iterate when posting a project, and the platform could generate rubrics that include prompts for the categories of feedback that relate to iteration. The interface for entering the feedback could also be accompanied by expert examples (e.g. see [94]) or in-situ guidelines that would help community members craft the feedback that is most useful for iteration. Configurable rubrics could also enable a designer to direct providers’ attention to where feedback is most needed.

We found that more responses, longer responses, and a lower proportion of positively-worded feedback correlate with iteration. Our results also showed that designer engagement with the feedback providers also correlates with iteration. More content, especially if distributed as previously described, and lower proportions of positive tone may also contribute to increased uncertainty about a design and prompt iteration. Additional responses could also signal community interest in the work, an implicit form of feedback that may also facilitate iteration.

Prior work has studied the attributes of design feedback received online that correlate with perceived quality [13]. The authors found that longer feedback, feedback with strong positive or negative tone, feedback with high language specificity, and feedback that asks questions or makes suggestions correlate with higher perceived quality. Our results showed that several of these variables – including feedback length, more critical tone, and thought-provoking statements – also correlate with iteration. It is therefore possible that perceived quality might be serving as a latent variable between the observed variables and iteration.

Our results showed that only a single design was posted for 79% of the projects in the forum data set; with two or more iterations being shared for 21% of the projects. These results were supported by the structured survey where a majority of the respondents reported posting only one design. The survey results indicated that designers are aware of the need to iterate and suggest that the observed rate of iteration is due in part to designers posting their designs near the end of their process, thus leaving insufficient time or desire to further iterate. This finding points to the need for more research to understand how to encourage designers to share their work earlier and more often. This is a potential issue not only for online design communities, but also for the crowd-based platforms that generate design feedback (e.g [7, 95, 96]).

The system implementation of the community studied did not directly support iteration. Consequently, designers have had to learn to re-appropriate the site’s mechanisms to represent iteration. For instance, a large majority (85%) of designers who iterated replied to their original thread with a comment linking to their revised design. The benefit is that this approach maintains a feedback history to contextualize subsequent discussion. However, the site’s presentation algorithm does not prioritize threads based on comment activity, even

if performed by the original poster. Revised designs therefore receive less attention than the initial project post. This outcome is consistent with the platform’s mission to promote discussion around fresh content rather than prolonged discussions around older content. A compromise could be for the site’s presentation algorithm to consider an iteration (e.g. signaled by the designer replying to their original post) as fresh content. It could also allow users to sort threads by the activity of the original poster.

The karma feature in the community may also have affected our results. Project posts that receive more upvotes gain more visibility on the landing page. This allows the project to attract more feedback, which facilitates iteration. However, in our data set, we found only a very weak correlation between upvotes and the number of iterations posted ($r=0.08$). Despite the lack of correlation, it is still possible some designers posted their work with the main objective of receiving upvotes and public exposure rather than critique.

For the projects that did iterate, it was surprising that the Turk participants did not perceive differences in quality between the initial designs and final iterations, despite the perception of moderate revisions. In contrast, the rater with design experience did detect differences in quality. We do not believe these outcomes are necessarily incongruent. The improvement in design quality may have been subtle and required more expertise to differentiate. This pattern of results should serve as a reminder of the need to include evaluations from those with domain expertise in future studies that collect subjective ratings of design quality.

3.6 LIMITATIONS AND FUTURE WORK

The data was extracted from a single platform in two month-long windows of time. The community studied (Reddit) was chosen primarily because of its active community and the ability to access the data through an API. Future work is needed to generalize the findings to how designers iterate over longer periods of time and in other online communities. Because the community studied does not explicitly make available a history of edits to the posts or comments, it is likely the heuristics used on our data set missed some iterations and identified some posts as iterations that were not. Given the size of the data set, we believe these points to be inconsequential to the findings of the work. Finally, we were unable to measure and include the expertise of the designers in our quantitative analysis; further research is needed to tease apart differences in iterative behavior between novice and expert designers in online communities.

We see three additional directions for future work. First, our results showed that even though designers could revise their work using the feedback from Reddit, the changes did not produce detectable improvements in the eyes of non-experts. Future work could compare how

designers motivated by their own goals would revise their work in response to expert feedback. A second direction would be to modify an online community to test the findings of this work. For example, one could incorporate rubrics or guidelines that promote the type of feedback that we found prompts iteration. Designers could also be allowed to organize iterations around their projects and showcase their process, rather than only individual designs. Last, future work could explore platform designs that address some of the psychological factors such as evaluation apprehension that may be deterring designers from sharing earlier versions of their work online.

3.7 CONTRIBUTIONS

The goal of this study was to learn how designers approach design iteration and seek feedback in online communities. Designers typically shared their projects with the community near the end of their process, with the expectation they would receive personalized, high quality feedback from a large, diverse audience. Designers who posted multiple iterations of their work typically received more process-oriented, comparison, and investigation of feedback. These types of feedback are often open-ended, questioning, and thought-provoking, creating uncertainty about a design's quality and helping designers recognize further revision and feedback are needed. The findings of this experiment indicate the quality, quantity, and types of feedback a designer receives can all affect how often they share iterations of their work online.

A paper reporting the results of this study was published at Creativity and Cognition 2017. The results described in this chapter contribute deeper empirical understanding of the characteristics of feedback received online that relate to design iteration, and how designers incorporate the use of online communities into their creative projects. These results can also inform how online communities and crowd-based design services (e.g., [7, 8]) can better represent projects and generate the types of feedback that promote iteration. Together, these contributions may enable online communities to serve as more effective resources for creators of open-ended design projects.

Data collected during this study revealed that designers typically approached the community for feedback a single time near the end of their design processes. During interviews, designers reported mixed results regarding the amount and quality of feedback they received, and cited additional factors they considered when soliciting feedback from the community. Such factors included the forum structure of the feedback and the ability to engage anonymously with their audience, both of which helped facilitate authentic discussions with a large, diverse user base of feedback providers. These findings inspired me to explore how the user interfaces present

throughout the feedback exchange process influence creators' perceptions, interpretation, and usage of the feedback they receive.

Chapter 4: Determining How Scores Mediate Interpretation of Written Comments

The preceding study revealed that presenting feedback in anonymous discussion threads had a significant impact on designers' decisions to share their work with an online community, inspiring me to investigate how user interfaces more broadly impact the way a creator perceives, interprets, and uses feedback. In addition to the constructive feedback sought by creators, the community discussed in the prior chapter leveraged upvotes to organize feedback within individual design posts. These votes are a form of summative feedback, which is primarily used to validate, affirm, and measure progress. Summative feedback in the form of grades or scores is often paired with constructive feedback, yet prior work reports mixed results regarding whether summative feedback is beneficial for open-ended works [2, 4, 19]. Additionally, little research has investigated how summative feedback influences creators' interpretation and usage of constructive feedback beyond its effects on task performance. In this chapter, I address these gaps by exploring how presenting various combinations of constructive and summative feedback influence a creator's perceptions of the feedback in terms of its helpfulness and fairness, and how these perceptions translate to revision depth, effort, and quality. This exploration takes the form of a quantitative study investigating how creators authored and revised short stories in response to a quality score, pre-authored constructive comments, both comments and a score, or no feedback.

4.1 INTRODUCTION

Constructive feedback is essential for helping content creators learn skills and improve outcomes for open-ended projects. In both online communities [10, 32, 53, 55, 97] and courses with large enrollments [15, 54, 67, 98], the growing demand for personalized constructive feedback has been outpacing the ability to generate that feedback, necessitating the development of automated feedback solutions. Current automated solutions typically map summative measures of quality (such as scores or grades) along dimensions of interest to pre-authored constructive written comments [69, 70, 71, 72, 99]. Despite the prevalence of this pattern, existing research is divided on whether assigning scores to open-ended projects is beneficial [2, 4, 5, 19, 100, 101]. Additionally, it is not well understood how scores give context to constructive feedback and influence subsequent revisions performed to open-ended work. For instance, does showing a score increase satisfaction with constructive feedback? Does showing a score increase the likelihood a recipient addresses constructive feedback? And do scores elicit higher revision effort or revision quality?

In addition to whether scores are presented or not, the magnitude of these scores may also influence feedback interpretation and subsequent revisions. High scores may maintain continued interest in pursuing a task [4] and reward strong performances [102], but may also downplay the perceived need to improve [2] and cause self-doubt in one’s ability to repeat strong performances [5]. Low scores may signal room for improvement [2] and motivate subsequent revision [3], but may also cause feelings of inadequacy [5] and discourage future participation in a task [4]. While prior work documents these effects of scores independently from comments, it remains an open question how high and low scores influence the interpretation and usage of constructive feedback on open-ended projects.

To address these gaps in the literature, we conducted an online experiment in which participants (N=441) wrote and revised short stories when presented with quality scores, pre-authored constructive comments, both scores and comments, or no feedback. We analyzed data from the story-writing task and two surveys to determine how this feedback affected participants’ task and feedback satisfaction, revision depth and effort, and overall improvement. We found the presentation of comments was correlated with higher feedback satisfaction, higher revision effort, and more improvement. We also found performance was positively correlated with task satisfaction among participants who were shown scores. Finally, we found showing either type of feedback elicited more deep revisions than showing no feedback, but showing constructive comments without scores elicited the most deep revisions.

Our results suggest constructive comments on open-ended work promote the most positive perceptions of feedback and most effective revisions when presented without explicit scores or grades. For those who receive high quality scores on their work, scores may elicit a sense of accomplishment and encourage continued pursuit of their open-ended work. However, scores may otherwise add limited value and context to constructive feedback on open-ended work, and can even undermine perceptions of constructive feedback and subsequent revisions performed. Designers of automated assessment platforms may incorporate these insights into alternative feedback presentations such as improvement scores or prioritized feedback lists to help learners and creators achieve desired revision outcomes.

Our primary contributions to the HCI community are 1) a deeper understanding of how scores and written comments affect creators’ perceptions of feedback and revisions to their open-ended work, and 2) practical design implications for automated assessment platforms aiming to effectively leverage the presentation of constructive feedback to learners and creators.

4.2 RESEARCH QUESTION AND HYPOTHESES

In this study, we address the question of how quality scores and pre-authored written comments influence creators’ perceptions of feedback and subsequent revisions to their creative writing. This question was posed to investigate how summative and constructive feedback interact to affect the process of revision in the context of automated assessments of open-ended work.

Creators may perceive scores alone as unfair and unhelpful oversimplifications of achievement towards their goals [3]. Constructive comments address this issue by allowing feedback providers to identify and justify potential solutions to problems within open-ended work [58]. This justification reduces a potential source of frustration and enriches the creator’s learning process. Our first hypothesis is that *(H1) receiving written comments with or without a score prompts higher task and feedback satisfaction than receiving only a score.*

Constructive comments help reveal gaps between how a creator conceptualizes their work and how others perceive it [87]. Reading comments enables creators to make deep, thoughtful revisions by reducing discrepancies between their understanding of a goal and performance towards that goal [1, 58]. Our second hypothesis is that *(H2) receiving written comments with or without a score prompts more deep revisions and greater improvement than not receiving comments.*

By quantifying the quality of their work, scores help creators compare themselves to their peers, devise strategies to improve, and optimize continued effort towards their goals [3]. Our third hypothesis is that *(H3) receiving a score with or without written comments prompts more revision effort and greater improvement than not receiving a score.*

One thread of prior work suggests summative assessments may aid creators in interpreting written comments [5, 100, 101]. However, another thread of research suggests showing scores on open-ended work may undermine the benefits of written comments [2, 4, 19], particularly if creators perceive the scores to be low. To investigate this discrepancy, we account for level of performance as a covariate in all of our analyses.

4.3 METHODOLOGY

4.3.1 Experimental Design

To test our hypotheses, we conducted a two-factor between-subjects study of how participants authored and revised short stories in response to feedback as part of a creative writing task. The factors were 1) showing vs. not showing pre-authored written comments and 2)

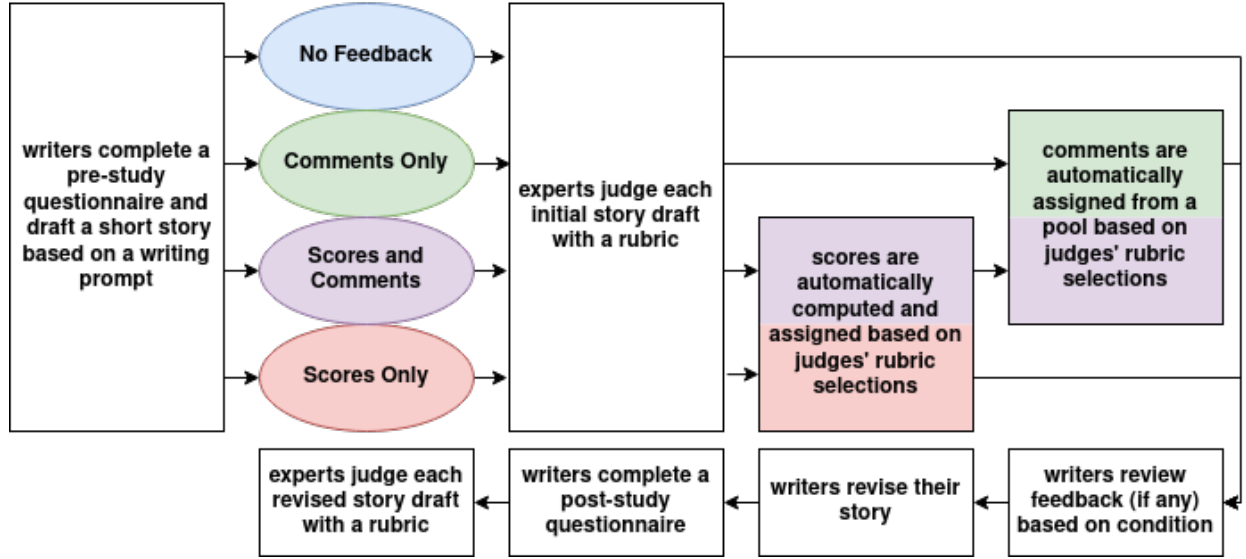


Figure 4.1: Methodology for the study

showing vs. not showing a quality score (see Table 4.1) in the feedback each participant received on their initial story drafts. We chose these factors as representatives of constructive and summative feedback typical of systems for automatically assessing open-ended work. We also examined how level of performance interacted as a covariate with scores and comments to influence participants' feedback perceptions and revisions to their stories.

We chose short story writing as a medium for our study because it exemplifies an accessible creative exercise incorporating both feedback and revision. Rather than fabricating scores and using prepackaged tasks, we elected to have participants write their own original stories and to provide them with authentic scores on their drafts. Given our sample size, we believed these choices would allow for a wide distribution of initial performance scores while increasing participants' sense of ownership over and investment in their work, lending ecological validity to the experiment and aligning with prior work [4, 18, 19, 44]. The wide range of scores also allows us to analyze the effects of initial performance on our dependent measures.

The final experimental design was informed by a series of pilot studies conducted on Amazon Mechanical Turk. A total of 27 participants took part in the pilot studies; those who participated in any of the pilot studies were not allowed to participate in the final study.

	score hidden	score shown
comments hidden	control	score
comments shown	comments	combined

Table 4.1: Experimental Conditions

Through the pilot studies, we were able to revise the task instructions for clarity and brevity, identify and correct usability problems with the task interface, and instrument effective data collection, among other improvements.

4.3.2 Participants

We recruited participants for our study through Amazon Mechanical Turk [103]. The platform was chosen due to prior success using it for large-scale creative writing tasks [104, 105, 106, 107] and infrastructure conducive to collecting and measuring written data.

We approached Reddit’s /r/mturk community [108] to solicit advice for recruiting participants from Mechanical Turk. At the community’s suggestion, we filtered participants to those with at least a 98% HIT (Human Intelligence Task) approval rate and 1000 HITs completed. These filters were implemented to ensure high quality data, minimize low effort work, and reduce the odds of participants attempting to game the task. We did not exclude participation based on past creative writing experience, though we did include story quality as a covariate in our analyses. To reduce potential language barriers, we limited participants to those residing within the United States. Additionally, to comply with IRB regulations, we restricted participants to those over 18 years of age. To incentivize high quality work, we offered participants \$4.00 base pay, and advertised a bonus of up to \$1.00 based on final story quality (e.g., a story receiving 75% of the max score would earn a \$0.75 bonus). In practice, everyone who completed the study was awarded the full bonus regardless of story quality.

We received submissions from a total of 526 participants across the 24-hour period the writing phase of our study was active. Of these, 17 submissions were discarded for plagiarism, while 68 participants did not complete the revision phase, leaving a total of 441 participants (245 female) who completed the study.

Participants ranged from 19-83 years of age, with the median age being 37 years. The vast majority of participants (85%) were native English speakers. About 39% of participants reported in engaging in creative writing infrequently, 21% monthly, 19% almost never, 17% weekly, and only 4% daily. However, most participants reported enjoying creative writing ($\mu=4.97$ out of 7, $\sigma=1.62$).

4.3.3 Task Design

Our study consisted of one primary story-writing task, split into two phases for writing and revision respectively. In the writing phase, each participant was asked to spend 10-15 minutes composing a short story between 125 and 250 words based on a visual writing prompt.


Revise Your Story

Welcome back! Your story has been reviewed by the judge panel and has received the following evaluation:

Score: **75/100**

Feedback: **The plot and setting are reasonably well-established; however, the story can be further improved by including more details about the plot or setting in which the story takes place, or providing more context for the opening or ending of the story. The characters are somewhat simplistic and under-developed, so revealing more about them by communicating their thoughts, appearance, action, or dialogue would better captivate readers. The story makes little use of physical and visual language, so there is plenty of room to enhance the story's quality by incorporating more lively language that appeals to the senses and vivid descriptions of the characters and their environment. Please note that while each of the components are graded separately, they work together, and that directed improvement toward one area may indirectly improve another. This feedback is meant to give you an idea of the areas in most need of general improvement.**

Prompt



Write a short story based on the image above.

Please write your story in the space below:

Harry let out a long sigh, turning back to look into the suitcase again. A small, pink unicorn toy, about a quarter of Harry's size, was pacing back and forward on the bottom of the suitcase, intermittently trying to leap out. Harry leaned down, scooped the unicorn up in one paw, and deposited him in the dirt beside the suitcase.

"Come on, man!" the unicorn explained in a deep tenor. "Can you believe this happened again, man? How many times is this going to happen? First Frankie ditches us, then Tommy, and

Current word count: **247 words**

Submit Story

Figure 4.2: Task interface for the story revision phase, as seen by a participant in the “combined” feedback condition. The interface for the initial writing phase was similar.

During pilot testing, we determined this range allowed for sufficient creative expression while enabling participants to finish within the suggested 10-15 minutes. In the revision phase, each participant was presented with written feedback, a numeric score, both, or neither depending on their condition, and was asked to revise their story as they saw fit. The second phase was otherwise identical to the first.

The task was administered through a self-contained web page created specifically for the study. In both phases, the page consisted of a writing prompt, a set of instructions, and a text field for composing a story. For the revision phase, the page also included any feedback the participants received, revision instructions, and a pre-filled form containing the participants' initial story drafts (Figure 4.2).

4.3.4 Data Collection and Metrics

The study had three measures: a pre-task background survey, the story-writing task itself, and a post-task questionnaire. The background survey was administered through an online form, and consisted of 12 questions. In addition to basic demographic information (age, gender, etc.), participants were asked several multiple choice and multiple answer questions about their experiences with creative writing (e.g., "How often do you engage in creative writing?") and receiving feedback (e.g., "In which context(s) have you received feedback on creative work?")

Within the task web page, story content was collected in a SQLite database using custom Javascript code. For each participant, story edits were tracked at a granularity of 15 seconds. The page also tracked when participants started and completed each phase of the writing task, as measured from typing their first character to submitting their story.

The questionnaire contained 17 questions about the time and effort participants put into their story drafts, the changes they made between drafts, the scores they believed they deserved on each draft, and their satisfaction with both the feedback they received and the task as a whole. The final set of items on the questionnaire were iteratively refined for brevity, clarity, and validity based on participant feedback during pilot testing. The questions consisted of 7-point agree-disagree Likert questions (e.g., "The feedback I received motivated me to revise my story"), short answer questions (e.g., "Approximately how much time did you spend revising your story?"), and multiple answer questions (e.g., "Please characterize the types of revisions that you made to your story"). Participants were also allowed to opt out of having their stories published in any research papers, and to enter any comments or questions they had with regards to the study or to the feedback they received.

4.3.5 Experimental Procedure

In the writing phase, after reviewing consent information, participants were asked to complete the background survey and create a temporary account on the study website. They were then directed to the task web page, where they were instructed to compose a short story based on a writing prompt (Figure 4.2). Participants were instructed to compose their story entirely within the research platform, and not to reuse an old story of their own or anyone else’s. Participants were notified that upon submission, their story would be reviewed by a judge with significant expertise in creative writing evaluation, that they would have a chance to revise and resubmit their story after the evaluation period, and that they would receive a bonus of up to \$1.00 based on the quality of their revised story. After submitting their stories, participants were informed they would receive feedback within 48-72 hours.

In the evaluation phase of the study, a judge reviewed each participant’s story and scored it according to a rubric. The judges consisted of two independently recruited graduate students from our university, both of whom had significant experience in formal writing critique. The rubric was developed by [109] and iteratively refined for the study by the research team and judges. Stories were evaluated along six dimensions: narrative cohesion, plot development, character exposition, sentence structure variation, writing mechanics, and adherence to instructions. Writing mechanics and variety in sentence structure were scored as either 1 (inadequate) or 2 (adequate), while all other dimensions were scored on a scale from 1 (needs improvement) to 4 (exceptional), for a maximum total of 20 points. Scores on individual dimensions were summed to form a final composite score for each participant.

Stories were scored by the judges across all dimensions besides “adherence to instructions”, which were scored automatically using a script. To ensure inter-rater reliability, the judges were calibrated on 9 sample stories (45 fields) with detailed instructions for using the rubric, and Cohen’s kappa was computed using quadratic weights [110, 111]. The weighted kappa was 0.63, signifying satisfactory agreement [112] for continuing with the review process.

For each dimension-score combination in the rubric, the research team and judges collaboratively developed a corresponding piece of pre-authored feedback. To minimize variation in the content of feedback, each piece was carefully constructed to use parallel language with respect to other feedback corresponding to the same score and dimension of evaluation (see Table 4.2 for examples). The feedback was also crafted to be justified, specific, actionable, and task-directed [57, 59], and of roughly uniform length, structure, and valence. Using this pool of feedback, a script automatically assigned three pieces of feedback to each participant based on the judges’ scores, prioritizing feedback on the weaker aspects of each participant’s work. This methodology for automatically assigning pre-authored constructive feedback based on

Dimension	Score	Feedback
Narration	3/4	The plot and setting are reasonably well-established; however, the story can be further improved by including more details about the plot or setting in which the story takes place, or providing more context for the opening or ending of the story.
Narration	1/4	The plot and setting are not well-established; the story is missing key information about the plot or setting in which the story takes place, or does not provide context for the opening or ending of the story.
Imagery	3/4	The story makes good use of physical and visual language, but incorporating additional lively language that appeals to the senses and vivid descriptions of the characters and their environment could enhance the story’s quality further.
Imagery	1/4	The story makes almost no use of physical or visual language; incorporating lively language that appeals to the senses and vivid descriptions of the characters and their environment are essential to enhancing the story’s quality.

Table 4.2: Examples of pre-authored constructive feedback given to participants. For each rubric dimension and score, the research team and judges developed a corresponding piece of feedback.

scores was chosen to mirror existing automated assessment systems [69, 70, 71, 72, 99].

Following the critiquing period, participants’ submissions were assigned randomly to one of the four experimental conditions. We controlled for initial level of performance during the assignment process to ensure approximately equal distributions of initial performance scores within each condition. In the “comments” condition, participants were presented with pre-authored constructive comments constructed from three items in the feedback pool based on the judge’s scores on individual dimensions of their work. In the “score” condition, participants were presented with a single composite quality score computed by adding the scores on each dimension of their work and converting to a 100-point scale. In the “combined” condition, participants were presented with both pre-authored constructive comments and a quality score as described for the “comments” and “score” conditions respectively. In the “control” condition, participants were presented with neither a score nor comments.

During the revision phase, participants in the “comments,” “score,” and “combined” conditions were asked to review the feedback they received and revise their stories with respect to that feedback. Those in the “control” condition were simply notified that their story had been reviewed, and were asked to revise their stories to the best of their ability. After submitting their revised stories, participants were instructed to complete the post-task questionnaire. Following the revision phase, each revised story draft was evaluated by the

Predictors	Block 1	Block 2	Block 3
Performance	0.047	0.047	-0.094
Score		0.250	0.247
Comments		0.238	0.235
Score : Cmts.		-0.361	-0.353
Perf. : Score			0.196**
Perf. : Cmts.			0.127
P : S : C			-0.096
Adj. R^2	0.006	0.007	0.028
Adj. ΔR^2	0.006	0.001	0.021
ΔF	3.686	1.154	4.129**

** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Table 4.3: Predictors of Task Satisfaction

Predictors	Block 1	Block 2	Block 3
Performance	0.082*	0.081*	0.085
Score		-0.159	-0.164
Comments		0.555*	0.553*
Score : Cmts.		N/A	N/A
Perf. : Score			0.092
Perf. : Cmts.			-0.099
P : S : C			N/A
Adj. R^2	0.009	0.038	0.043
Adj. ΔR^2	0.009	0.029	0.005
ΔF	4.098*	5.982**	1.922

** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Table 4.4: Predictors of Feedback Satisfaction

same judge who reviewed that story’s initial draft.

4.4 RESULTS

Out of a 20-point scale, scores for initial story drafts ranged from 9-20 points ($\mu=14.10, \sigma=2.23$), while scores for revised story drafts ranged from 8-19 points ($\mu=14.00, \sigma=2.26$). For their revised stories, 58% of participants revised their sentence structure, 58% revised their plot, 41% revised their use of imagery, 40% revised their character development, 35% revised their spelling and grammar, 33% revised their story length, and 4% did not perform any revisions.

We conducted hierarchical linear regressions with three blocks each on our entire sample

Predictors	Block 1	Block 2	Block 3
Performance	0.022	0.023	0.017
Score		0.317*	0.316*
Comments		1.210***	1.211***
Score : Cmts.		-0.358*	-0.352*
Perf. : Score			-0.053
Perf. : Cmts.			-0.004
P : S : C			0.138
Adj. R^2	0.000	0.238	0.243
Adj. ΔR^2	0.000	0.238	0.005
ΔF	0.955	47.002***	1.988

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.5: Predictors of Deep Revisions

Predictors	Block 1	Block 2	Block 3
Performance	0.037	0.037	0.010
Score		0.100	0.100
Comments		-0.280*	-0.281*
Score : Cmts.		-0.106	-0.105
Perf. : Score			0.039
Perf. : Cmts.			0.039
P : S : C			-0.048
Adj. R^2	0.005	0.031	0.026
Adj. ΔR^2	0.005	0.026	-0.005
ΔF	3.351	4.835**	0.201

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.6: Predictors of Shallow Revisions

of participants who completed the study ($n=441$). Hierarchical linear regressions were chosen to isolate the effects of presenting a score or written comments from the effects of participants' levels of performance. After verifying there was no considerable correlation among our dependent variables, we decided to use several independent regressions instead of a single multivariate regression to better reflect the structure of our research hypotheses. Because F-tests at each block of our regressions account for multiple comparisons within each model, and because our research hypotheses make specific predictions about significant effects we expect to find between models, we do not perform any extra adjustments for multiple comparisons [113, 114, 115]. Per convention, we report all comparisons in Tables 4.3-4.8.

Six hierarchical regressions were conducted in total, using task satisfaction, feedback

Predictors	Block 1	Block 2	Block 3
Performance	-1.000***	-1.000***	-1.451**
Score		-0.079	-0.100
Comments		6.192***	6.150***
Score : Cmts.		-1.512	-1.422
Perf. : Score			-0.162
Perf. : Cmts.			0.975
P : S : C			0.172
Adj. R^2	0.030	0.074	0.077
Adj. ΔR^2	0.030	0.044	0.003
ΔF	14.440***	7.976***	1.449

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.7: Predictors of Revision Effort

Predictors	Block 1	Block 2	Block 3
Performance	-0.146***	-0.146***	-0.158**
Score		0.118	0.118
Comments		0.673***	0.672***
Score : Cmts.		-0.186	-0.185
Perf. : Score			0.016
Perf. : Cmts.			0.006
P : S : C			0.001
Adj. R^2	0.063	0.110	0.105
Adj. ΔR^2	0.063	0.047	-0.006
ΔF	30.720***	8.692***	0.038

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.8: Predictors of Improvement

satisfaction, number of deep revisions, number of surface revisions, revision effort, and revision improvement as the respective dependent variables in each regression. The 107 participants in our control condition were excluded from our analysis of feedback satisfaction (n=334) due to the corresponding survey question being irrelevant. We otherwise included all participants in each of our remaining analyses (n=441). We define the metrics for our dependent variables and elaborate on our decisions when discussing the individual regressions. For each regression, the first block consisted of level of performance on the initial story draft (i.e., the quality score out of 20 points) as the sole predictor. The second block added the experimental conditions (i.e., presence of score, presence of comments, and their interaction) as predictors. The third block added the interactions between level of performance and

experimental conditions as predictors.

We used the statistical software package R to perform all regression analyses. Diagnostic plots and tests for each regression indicated that the assumptions of homoscedasticity, linearity of the data, normality of residuals, and non-multicollinearity were all met.

4.4.1 Creator Dispositions (Task and Feedback Satisfaction)

We measured task satisfaction using participants' agreement with the following 7-point Likert prompt on the questionnaire: *I enjoyed the overall experience of writing and revising my story*. For task satisfaction, the first block of our model was not significant, with initial performance explaining 0.6% of the variance. The second block of our model explained 0.1% additional variance and was not significantly better than the first. The third block of our model explained an additional 2.1% of the variance and was significant ($\Delta F(3, 433)=4.1289, p=0.007$). There was a significant interaction effect between level of performance and presence of a score ($\beta=0.19553, t=2.821, p=0.005$), but not between level of performance and presence of comments. These results do not support our hypothesis (H1) that comments prompt more task satisfaction than scores. The interaction between level of performance and presence of a score suggests receiving high scores increase task satisfaction.

We measured feedback satisfaction using participants' agreement with the following 7-point Likert prompt on the questionnaire: *The feedback I received reflected the quality of my initial story*. Because this prompt did not make sense for participants who received no feedback, we excluded participants in the control condition ($n=107$) from our analyses. For feedback satisfaction, the first block of our model was significant ($F(1, 332)=4.098, p=0.044$), with initial performance ($\beta=0.08211, t=2.024, p=0.044$) explaining 0.9% of the variance. The second block of our model was significantly better than the first ($\Delta F(2, 330)=5.982, p=0.003$) and explained 2.9% additional variance. Only presence of comments ($\beta=0.55492, t=2.513, p=0.013$) had significant effects on feedback satisfaction. The third block of our model was not a significant improvement over the second, explaining only 0.5% additional variance. These results support our hypothesis (H1) that comments prompt higher feedback satisfaction than scores. The modest positive correlation between performance and feedback satisfaction may indicate participants who believed they performed well were more receptive towards constructive criticism.

Prior work on expectation violation has shown receiving a lower score than expected may increase the score recipient's attentiveness towards written feedback [116]. In our questionnaire, we asked participants "On a scale from 0 to 100 (highest score), what score do you think you deserved for your initial story?" Based on responses to the above question, we

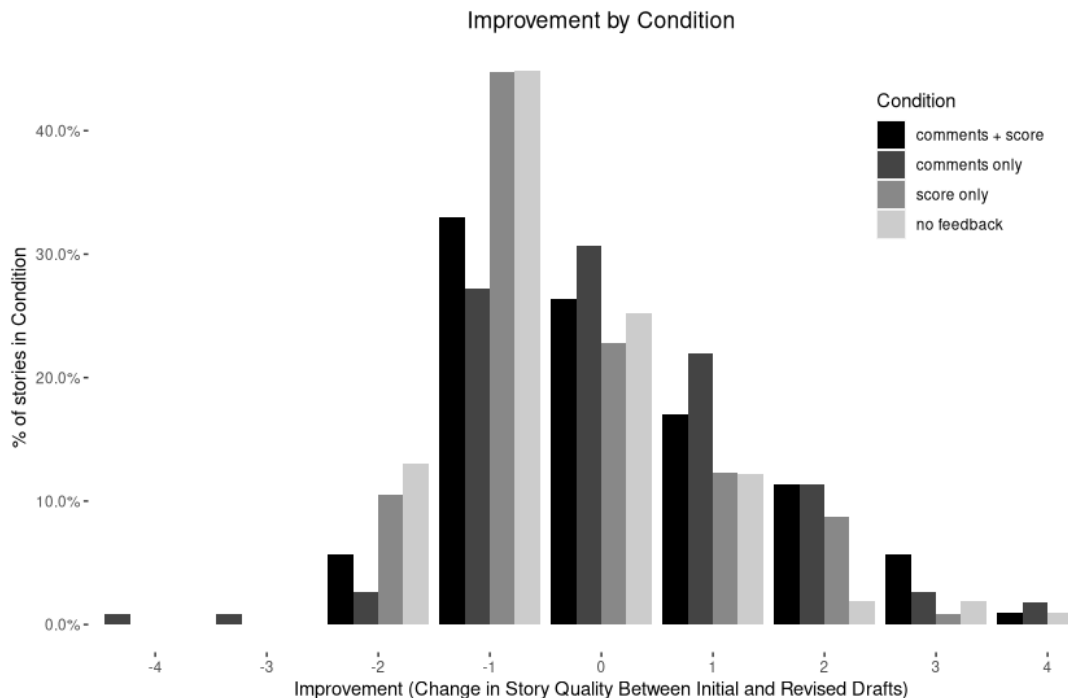


Figure 4.3: Distribution of improvement scores by condition, computed as the difference between the quality scores for revised and initial story drafts. Participants who received comments typically received higher quality scores than those who did not receive comments.

partitioned participants who were shown scores into two groups: those who received a score as least as high as they felt they deserved, and those who received a score lower than they felt they deserved. However, post-hoc T-tests between these groups revealed no significant effects on any outcome variables.

4.4.2 Revision Behavior (Depth of Revisions)

Participants were asked about the types of revisions they performed using the following questionnaire prompt: *Please characterize the types of revisions that you made to your story.* While per-dimension score differences between participants’ initial and revised submissions were another candidate for measuring revision behavior, we chose to use self-reported revisions to account for the possibility of revision without improvement. We allowed participants to select as many items as they liked from a list of six specific revisions, which we grouped into two categories. Revisions to plot, character development, or use of imagery were classified as “deep” revisions, while revisions to spelling and grammar, sentence structure, or story length were classified as “surface” revisions. Although our list offered an “other” option, only 6% of participants used it, usually to clarify their selection of other revisions. We interpreted this to

mean our revision list was sufficiently comprehensive, and we excluded these other revisions from our analyses. Based on responses to the survey prompt, we counted the number of distinct types of deep and surface revisions participants made, and used these counts for our regression analyses.

For deep revisions, the first block of our model was not significant and explained no variance. The second block of our model was significantly better than the first ($\Delta F(3, 436)=47.002, p < 0.001$) and explained 23.8% of the variance. Both presence of scores ($\beta=0.31701, t=2.524, p=0.01$) and presence of comments ($\beta=1.21014, t=9.637, p < 0.001$) had significant effects on the number of deep revisions made, signifying both are effective at eliciting substantial revisions, albeit comments more so. However, the interaction between presenting scores and comments was also significant ($\beta=-0.35793, t=-2.013, p=0.04$). The negative effect size in this interaction suggests that while scores may prompt more deep revisions than no feedback, presenting both comments and a score is no more effective than presenting comments alone. The third block of our model explained only 0.5% additional variances and did not significantly improve the model. Overall, these results support our hypothesis (H2) that receiving comments prompts more deep revisions than not receiving comments.

For surface revisions, the first block of our model was not significant, with initial performance explaining only 0.5% of the variance. The second block of our model was significantly better than the first ($\Delta F(3, 436)=4.835, p=0.003$) and explained 2.6% additional variance. Only presence of comments had a significant negative correlation with the number of surface revisions made ($\beta=-0.280, t=-2.212, p=0.03$). The third block of our model was not significantly better than the second block and did not explain any additional variance. Though we did not directly hypothesize about surface revisions, the lower number of surface revisions among participants who received comments complements the higher number of deep revisions, lending additional support to H2.

4.4.3 Creative Outcomes (Effort and Improvement)

We measured revision effort as the difference between the time participants spent working on their revised stories and on their initial stories. This metric was chosen over absolute revision time to account for individual variation in effort. Using this metric, positive values indicated more time was spent on the revised story draft and negative values indicated more time was spent on the initial draft. For revision effort, the first block of our model was significant ($F(1, 439)=14.440, p < 0.001$), with initial performance ($\beta=-0.99780, t=-3.813, p < 0.001$) explaining 3.0% of the variance. The negative effect size suggests that participants who performed well initially were likely to invest less time revising later. The second block of

our model was significantly better than the first ($\Delta F(3, 436)=7.976, p < 0.001$) and explained 4.4% additional variance. Only presence of comments had a significant effect on revision effort ($\beta=6.192, t=3.850, p < 0.001$), suggesting written comments motivated recipients to invest more effort in performing revisions. The third block of our model was not a significant improvement over the second block and explained only 0.3% additional variance. These results do not support our hypothesis (H3) that receiving a score prompts more revision effort than not receiving a score, indicating other factors may be more important.

Improvement was computed for each participant as the difference between the quality scores for their revised and initial story drafts. Figure 4.3 shows the distribution of improvement scores for participants in each condition. For improvement, the first block of our model was significant ($F(1, 439)=30.720, p < 0.001$), with initial performance ($\beta=-0.1464, t=-5.543, p < 0.001$) explaining 6.3% of the variance. The negative correlation between initial performance and subsequent improvement may indicate a ceiling effect on potential improvement for those with already-high scores. The second block of our model was significantly better than the first ($\Delta F(3, 436)=8.692, p < 0.001$) and explained 4.7% additional variance. Only presence of a comments had a significant effect on improvement ($\beta=0.673, t=4.155, p < 0.001$), reaffirming the effectiveness of specific, actionable, task-directed feedback on creative work. The third block of our model was not significantly better than the second block and did not explain any additional variance. In general, these results support H2 but do not support H3 regarding the respective effects of comments and scores on improvement.

4.5 DISCUSSION AND FUTURE WORK

Participants who received constructive comments on their stories were more satisfied with their feedback than those who received scores, but were no more satisfied with their experience working on the writing task itself (H1). By contrast, participants' satisfaction with their story-writing experiences was positively correlated with the magnitude of their scores, regardless of whether they received comments. Although task satisfaction and enjoyment may incentivize continued participation in creative endeavors [4, 117, 118], the novices who would most benefit from this incentive may be less likely to receive high scores on their work. This presents a challenge to designers of automated assessment systems in framing negative evaluations without discouraging novice learners and creators. Rather than showing absolute quality scores, one potential design implication might be to show scores indicating how much a creator has improved their work since their last draft.

As hypothesized, participants who received written comments made more deep revisions and greater improvements to their work than those who did not receive comments (H2).

Though both comments and scores were positively correlated with more deep revisions, the effect size of comments was nearly four times larger, while scores and comments together elicited no more deep revisions than comments alone. A long history of work affirms the value of constructive comments in helping creators conceptualize problems with their work and make deep, thoughtful revisions [1, 13, 58, 87, 101, 119, 120]. The negative correlation between initial level of performance and revision improvement may reflect diminishing returns when trying to improve already high-quality work. In these scenarios, it is possible creators may not find deep revisions necessary or desirable for improving their open-ended projects. This does not necessarily mean comments should be withheld to encourage surface-level revisions, but rather suggests comments operating at a lower level of detail should be employed when deep revisions are unwarranted. Automated assessment systems might operationalize this by drawing from different feedback pools depending on a creator’s desired level of feedback detail or the assessed quality of their work.

Our hypothesis that receiving a score prompts more revision effort and greater improvement than not receiving a score (H3) was not supported. Scores had no significant effect on the effort participants put into their revisions, nor on how much those revisions subsequently improved their work. This result may reflect issues surfaced in one thread of research showing external rewards (including scores) can preclude effort by undermining learning aspects of a task [121] or by indicating little work remains to be done [122, 123]. Our findings suggest the utility of scores on open-ended projects may be more limited than that of constructive comments. Because mapping quality scores to constructive feedback is a common pattern among existing systems for automatically assessing open-ended work, designers of these systems ought to carefully consider whether presenting these scores is in creators’ best interests.

Our findings reaffirm prior research suggesting that constructive written comments are beneficial to creators [1, 5, 58, 87]. However, our findings also suggest that the presence of scores adds little additional value to critiques of open-ended work, and in some cases may undermine the benefits of constructive feedback. Given these findings and the prevalence of quality scores among current automated assessments of open-ended work, we propose several design alternatives to including quality scores within these assessments. One strategy would be to have automated assessment systems continue mapping scores to constructive feedback without presenting the scores themselves to creators. Another alternative might be to show scores indicating how much a creator has improved their project since the last time their work was assessed. This approach may be especially beneficial to novices, as it avoids the pitfall of quantifying a work’s absolute quality while framing the score in a positive light to incentivize continued work on the project. Scores associated with constructive feedback

could also be positively recontextualized as the importance of addressing that feedback. For instance, automated assessment systems could incentivize revision by implementing a scoring mechanism where creators earn more points on their work by performing more important revisions. Finally, the provision of a score could be left up to the creators themselves. As one example, when submitting their work for assessment, a creator could indicate whether they would prefer any constructive comments they receive to include improvement scores, importance scores, or no scores.

For our experiment, we selected pre-authored comments and quality scores as representative samples of constructive and summative feedback commonly used within automated assessment systems. Other forms of constructive and summative feedback such as oral critiques, social media upvotes, peer rankings, and graphical visualizations could have different impacts on creators’ perceptions of feedback and revisions to their work. Future work is needed to explore how alternate presentations of feedback influence revisions to open-ended work. We also chose to study creative writing because it typifies a class of domains suitable for automated assessment in which people iteratively revise open-ended work with respect to feedback on in-progress drafts. Additional work is needed to test how our results generalize to other open-ended domains such as graphic design, music composition, academic writing, and programming. Finally, the quantitative nature of our study invites subsequent research to investigate creators’ perceptions of scores and written comments from a qualitative perspective.

4.6 LIMITATIONS

On average, the quality of participants’ stories remained the same between initial and revised drafts (initial draft scores: $\mu = 14.10$ out of 20; revised draft scores: $\mu = 14.00$ out of 20). A post-hoc analysis revealed that the judges scored unchanged stories one point lower on average during the second evaluation period. While decision fatigue [124] and increased judge expectations for revised drafts may explain these trends, the absence of discernible improvement may also be attributable to a novice population. Novices have been shown to make revisions that, when evaluated by experts, were deemed no better than their original work [125], even if they perceive the feedback to be helpful and of high quality [7]. A follow-up study might compare and contrast how a more experienced population revises open-ended work in response to feedback containing scores and written comments.

To keep the experiment tractable, we limited stories to one revision cycle and 250 words after determining through pilot studies these constraints allowed for sufficiently well-developed stories. While some participants felt these constraints helped them revise more thoughtfully,

others felt limited in expressing their creativity and addressing the feedback they received. Future work should examine how scores and comments affect revision behavior across multiple unconstrained drafts.

We conducted our study on Amazon Mechanical Turk, a platform where users are primarily incentivized through financial compensation. Additional research is needed to generalize our results to participant pools driven by different motives such as social status or enjoyment.

4.7 CONTRIBUTIONS

Constructive and summative feedback are critical components of systems for automatically evaluating open-ended work, yet little existing research has investigated how these components affect creators' perceptions of feedback or the subsequent revisions they make. In this chapter, we explored how presenting pre-authored constructive comments, quality scores, both comments and scores, or no feedback influenced task and feedback satisfaction, revision depth and effort, and improvement to short stories. We found that the presentation of comments was positively correlated with feedback satisfaction, revision depth, revision effort, and improvement. While performance was positively correlated with task satisfaction when a score was present, scores did not otherwise have any significant benefits to recipients, and in some cases undermined the benefits of comments. We hope our results can inform the design of future automated assessment systems through insights regarding the judicious presentation of scores and constructive feedback to creators of open-ended projects.

This work makes two major contributions. First, it provides a deeper understanding of how scores interact with written comments to affect perceptions of feedback and subsequent revisions to open-ended work. The results suggest constructive comments on open-ended work promote the most positive feedback perceptions and effective revisions when presented without explicit scores. While scores may elicit a sense of accomplishment and encourage continued pursuit of open-ended work, they otherwise add limited value to constructive feedback on open-ended work, and may even undermine perceptions of the feedback and subsequent revisions performed. Second, it provides practical design implications for feedback exchange platforms aiming to effectively leverage the presentation of constructive feedback to creators. Designers of these platforms may incorporate insights from this study into alternative feedback presentations such as improvement scores or prioritized feedback lists to help creators achieve desired revision outcomes. A paper reporting the results of this study was published at Learning at Scale 2021.

The findings from this study revealed that showing quality scores on creative, open-ended works has minimal benefit compared to showing constructive written comments without

quality scores. While my experiment tightly controlled the length, tone, structure, and content of written comments to examine interaction effects with scores, these attributes vary greatly in real-world constructive feedback. This variance may stem from the feedback provider's expertise and writing style, but may also arise from the interface used to compose the feedback. Additionally, a creator's ability to interpret the feedback they receive may also depend on the interface used to present the feedback. These observations inspired my next study exploring how interfaces for composing and presenting feedback at different levels of detail influence the composition, interpretation, and usage of the feedback.

Chapter 5: Investigating How Feedback Detail Affects Feedback Composition and Interpretation

In the prior section, I explored how presenting a score with constructive comments can influence the way creators interpret and act upon those comments. Although my previous study concluded that constructive comments without quality scores are most effective for open-ended creative works, the study design did not address the real-world variance in how feedback is structured and presented. In this chapter, I investigate how presenting feedback organized at four different levels of detail influences creators' perceptions of the feedback, and how these perceptions translate to revisions and creative outcomes. Each level was chosen to represent a feedback pattern commonly used in academic and workplace environments: rubrics, open-ended comments, rubrics with open-ended comments, and rubric with comments on each rubric criterion. While various works have argued for the benefits of some of these patterns over others, little work has compared their relative merits side by side, and even less has examined these relative merits from both the provider's and recipient's perspective. Towards this end, I conduct a parallel investigation of how feedback providers weigh the perceived difficulty of composing feedback at each level of detail against their perceived value of that feedback to the recipient.

5.1 INTRODUCTION

The demand for feedback is growing due to instructors incorporating design-based learning processes into courses [15, 58, 101, 126, 127] and increased participation in online creative communities [128, 129, 130, 131, 132]. Producing feedback at scale requires balancing the needs of the feedback providers (e.g., low cost of feedback composition) with the needs of the recipients (e.g., receiving the most helpful feedback). From the recipient's perspective, feedback is most helpful when it is timely, justified, and task-directed [58, 59, 133]. From the provider's perspective, feedback should be inexpensive, consistent, and easily accessible [134, 135, 136, 137].

Although these ideals are not mutually exclusive, limited resources often necessitate compromises at one or more steps of feedback generation. The tradeoffs of producing feedback with one technique over another therefore necessitate careful consideration of which feedback attributes are most valuable to each party. Though inexpensive and accessible, feedback created with rubrics may be perceived by recipients as less fair and invested than personalized comments written in a open text field [138, 139], and providers may feel less capable of producing helpful feedback using rubrics alone [140, 141]. While scaffolding may

reduce the cognitive load required to compose feedback, scaffolded composition techniques such as per-criterion comments may hinder expressiveness and result in less personalized feedback [142, 143, 144, 145]. Recipients may also perceive and act on the feedback differently depending on whether the feedback comes from a peer or an expert [146]. Despite the unique tradeoffs for each of these techniques, little research has compared these tradeoffs from the perspectives of both feedback providers and recipients.

To explore these tradeoffs between composition techniques, we conducted a two-factor between-subjects experiment of how feedback presentation and perceived source influenced novice participants' feedback perceptions and revision outcomes. For a creative writing task, we presented participants ($N=285$) with feedback in the form of a writing rubric, open comments, a rubric with open comments, or a rubric with per-criterion comments. We explored how these increasingly detailed presentations influenced participants' perceived fairness, credibility, investment, and helpfulness of the feedback, as well as the extent to which participants were able to revise and improve their stories. We also analyzed how these measures were affected by whether the feedback was perceived to be from an experienced writer (expert) or another participant in the study (peer). Finally, we measured the time and perceived effort required by expert providers ($N=4$) to compose the feedback in each condition, and how these providers perceived the value of the feedback they composed.

We found that participants made non-trivial improvements to their work regardless of the type of feedback they received. Participants' revision quality and perceptions of feedback helpfulness, credibility, and investment increased along with the feedback's level of detail. Feedback providers stated rubrics with open comments would likely be as helpful to participants as rubrics with per-criterion comments, despite spending substantially more time composing the latter. On the other hand, composing feedback with rubrics alone required half the time of writing per-criterion comments, and the resulting feedback enabled participants to make improvements to their work comparable in quality to those who received open comments. Though providers reported dissatisfaction with the lack of expressivity when composing feedback using rubrics alone, providers also felt the extensive scaffolding of per-criterion comments diminished their focus and ability to compose effective feedback. We distilled these findings into recommendations for balancing provider and recipient needs, such as using per-criterion comments to maximize revision quality, open comments to maximize provider satisfaction, and rubrics to minimize composition costs.

Our work makes four main contributions to the HCI community. First, we contribute empirical knowledge of how the presentation of feedback influences a recipient's perceptions of the feedback and subsequent revision outcomes. Second, we offer insights linking these perceptions and outcomes with the feedback's composition cost and perceived value from

the provider’s perspective. Third, we identify provider perceptions surrounding the use of scaffolding in feedback composition interfaces, and how this scaffolding influences providers’ attitudes and processes towards feedback composition. Finally, we contribute an emergent framework for selecting feedback composition techniques based on specific attributes of interest, (such as revision quality or composition cost,) and insights relating the costs of composition to the benefits of the resulting feedback.

5.2 RESEARCH QUESTIONS AND HYPOTHESES

In this study, we explore how recipients revise their creative work in response to feedback composed with different levels of detail and perceived to be from providers with different expertise. We also explore the cost and perceived value of the feedback at corresponding levels of detail from the perspective of these feedback providers. Our exploration consists of three research questions:

RQ 1: *How does receiving feedback with different levels of detail affect the revisions performed on the work targeted by that feedback? How does the perceived expertise of the source of the feedback (e.g., peer or expert) mediate this effect? (recipient perspective)*

RQ 2: *How does receiving feedback composed with different levels of detail affect the recipient’s perceptions of the feedback (e.g., its helpfulness)? How does receiving the feedback from sources with different expertise mediate these perceptions? (recipient perspective)*

RQ 3: *How do feedback providers perceive the value of the feedback that they compose at different levels of detail? How does the level of detail of the feedback affect the time and effort needed to compose it? (provider perspective)*

We expect the cost of producing feedback to increase with feedback level of detail. However, we also expect recipients to perceive detailed feedback more favorably and to perform higher quality revisions in response to this feedback. In answering our research questions, we aim to link the costs of various feedback composition techniques with the benefits the resulting feedback confers to recipients. This empirical guidance will help instructors and designers of large feedback-driven communities make informed selections of feedback techniques to maximize outcomes of interest.

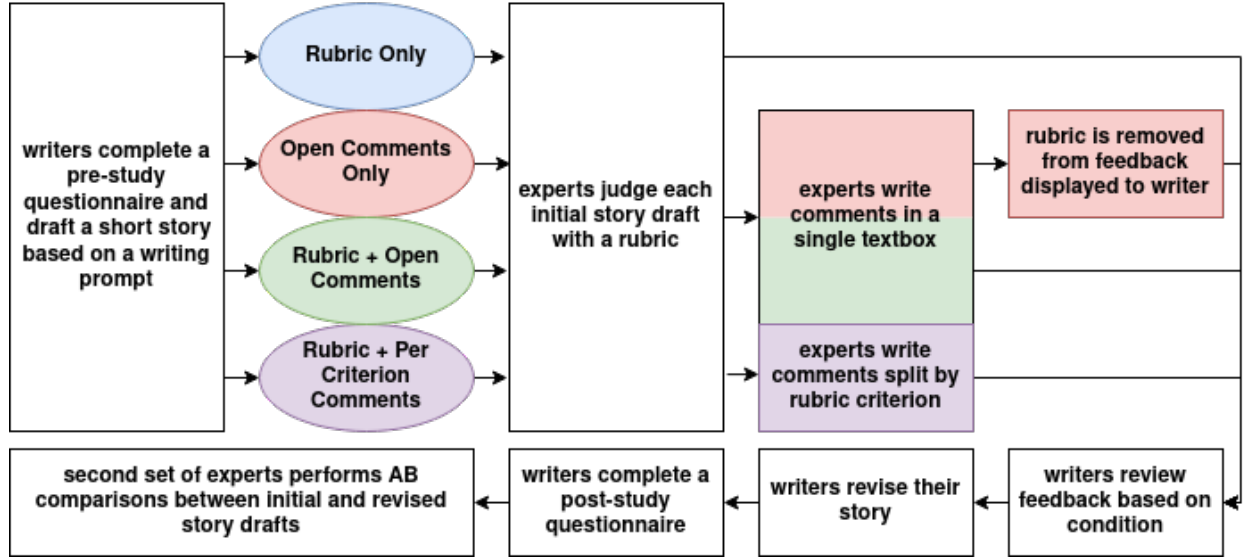


Figure 5.1: Methodology for the study

5.3 METHODOLOGY

To answer our research questions, we conducted a between-subjects experiment with two factors: *level of detail in the feedback* (in order of increasing detail: rubric only, open comments only, rubric with open comments, and rubric with per-criterion comments [scaffolding]) and *perceived source* (expert vs. peer). In the experiment, participants drafted short stories and revised the stories in response to feedback composed with different levels of detail and perceived to be from either domain experts or peers. We will use the term “participants” to refer to those who wrote and revised stories and “providers” to refer to those who composed the feedback for the participants.

Our experiment consisted of *writing* (participant), *evaluation* (provider), and *revision* (participant) phases. In the writing phase, participants drafted a creative story in response to a writing prompt. In the evaluation phase, providers composed feedback with a level of detail corresponding to participants’ assigned experimental conditions. In the revision phase, participants revised their stories based on the providers’ feedback. Half the participants in each condition were further informed their feedback was developed by a reviewer experienced in creative writing (expert source). The other half were informed their feedback was developed by another participant in the study (peer source).

5.3.1 Feedback Composition

Participants were presented with feedback at one of four levels of detail depending on their experimental condition:

- *Rubric only*: Participants in this condition received assessments of their stories on five criteria: narrative cohesion, character exposition, use of imagery, sentence structure variety, and writing mechanics. Such schemas combining high and low level narrative elements have been shown to effectively elicit high quality stories from crowd workers [147]. A provider marked the proficiency for each criterion: Exceptional, Above Average, Developing, or Needs Improvement. Each mark was supplemented by a pre-authored statement explaining it. Borrowed from [109], the rubric was refined for the experiment by the research team in collaboration with two graduate students in an English program at our university. Prior work shows rubrics enable non-experts to compose feedback comparable in quality to that of experts [13]. However, the limited detail of rubric-based feedback may limit the scope of revisions performed in response to this feedback. This experimental condition is also representative of the use of templates [61, 62, 63] and automated assessment [65, 148] where the recipient is assessed only on pre-defined criteria.
- *Open comments only*: In this condition, participants received a single text box of comments on the quality of their stories and suggestions for improving them. Thinking about and writing free-form comments likely requires more effort than using a rubric, but the personalized and unconstrained statements might also prompt more revision and favorable perceptions from a participant.
- *Rubric with open comments*: Participants received both a rubric assessment and open comments on their stories, as described in the first and second conditions respectively. While both rubrics and open comments may effectively encourage revisions, the combination of both may prompt deeper revisions than either presentation individually.
- *Rubric with per-criterion comments (scaffolding)*: In this condition, participants received the same rubric assessment and open comments as described in the prior condition. Participants also received additional comments explaining the marks and suggesting related improvements for narration, characterization, and imagery. Though likely requiring more effort to compose than the preceding conditions, detailed comments on individual rubric criteria might also induce the strongest revisions. Prior work has also suggested novices prefer feedback that is specific but comprehensive whenever possible [10].

Participants were balanced across these four conditions following the writing phase of the study. Because some participants did not return for the revision phase of the study, the final number of participants in each condition is imbalanced. Our chosen conditions are representative of several classes of composition techniques that incur different costs, and that are used in practice within online learning and design platforms. Although our selection does not exhaustively represent all possible techniques, it takes a preliminary step towards understanding how the composition costs and benefits of feedback are linked.

In addition to these participants, we recruited four feedback providers to evaluate participants’ initial stories. Providers composed feedback for each story using one of three user interfaces corresponding to the author’s feedback presentation condition. The layout of each composition interface mirrored the layout of the corresponding feedback presented to participants. Following a within-subjects design, each provider evaluated approximately the same number of stories using each interface, and each story was evaluated by exactly one provider. To minimize order effects on providers’ perceptions and evaluations, each provider was required to work through the three interfaces in a different order, composing all reviews with a given interface before proceeding to the next interface.

The open comments condition was implemented by having providers compose feedback using the same interface as in the rubric+open comments condition, but presenting only the comments to participants. This allowed us to use providers’ rubric selections to measure the quality of the initial stories in each condition. To ensure providers used each interface for approximately the same number of reviews, we split participants between the open comments and rubric+open comments conditions. We made these practical decisions to minimize the complexity of the experiment while maintaining a sample size sufficient for analyses in each condition. Finally, because numeric ratings may negatively influence feedback composition [12], we avoided explicitly showing providers numeric scores in the feedback interfaces, using only labels (e.g., “Exceptional”, “Needs Improvement”, etc.) Table 5.1 summarizes the number of participants in each experimental condition.

5.3.2 Participants and Feedback Providers

Story-writing participants (N=285) were recruited from Amazon’s Mechanical Turk [103]. We chose Mechanical Turk due to prior successes using it for large-scale experiments with creative writing tasks [104, 105, 106, 107] and infrastructure conducive to collecting and measuring written data. Prior work has also demonstrated Turkers are capable of writing creative stories [149] and evaluating and acting on the feedback received according to narrative structure [97].

Feedback Presentation	Rubric Only		Open Comments		Rubric + Open Comments		Rubric + Scaffolding	
Perceived Source	Expert	Peer	Expert	Peer	Expert	Peer	Expert	Peer
No. Participants	52	46	27	21	31	20	42	46

Table 5.1: Number of participants in the experimental conditions. Participants were presented with feedback in four conditions and were informed their feedback came from either an expert or peer. Four providers rotated through all the composition techniques in different orders to compose feedback for the participants. As it had no bearing on any outcome measure, perceived source was dropped from the final analyses. Imbalances result from splitting the two open-comments conditions and from some participants skipping the revision phase.

We filtered participants to those with at least a 99% HIT (Human Intelligence Task) approval rate and 1000 HITs completed. These filters were implemented to ensure high quality data, minimize low effort work, and reduce the odds of participants attempting to game the task. To reduce potential language barriers, we limited participants to those residing within the United States. Additionally, we restricted participants to those 18 years of age or older. To incentivize retention and high quality work, we offered participants \$5.00 for making a good faith effort to write and revise an original story, remunerating the first \$1.00 after participants submitted an initial story. These parameters were based in part on input that we sought from Reddit’s /r/mturk community [108]. The experiment was approved by our university’s IRB.

Participants completed an online survey asking six demographic questions (age, gender, English fluency, etc.). The survey also included two multiple choice questions about participants’ experience writing stories (“On average, how often do you engage in creative writing?”) and receiving feedback (“How often do you receive feedback on your creative writing?”) For writing experience, options included “less than once per month”, “about once per month”, “about once per week”, and “more than once per week”. For receiving feedback, options included “almost never”, “sometimes”, “usually”, and “almost always”.

Participants ranged from 21 to 82 years old (median = 37 years), and were predominantly native English speakers, with 87% reporting being “Perfectly Fluent.” About 49% of participants reported engaging in creative writing “less than once per month”, while 13% reported engaging in creative writing “more than once per week”. Similarly, around 88% reported receiving feedback on their writing “sometimes” or “almost never”, while 12% reported receiving feedback “usually” or “almost always.” The participant data indicates our sample was mostly novice writers with minimal experience acting upon feedback on their work, representative of novices receiving feedback in online creative communities or a first design-oriented course.

In addition to the participants writing the stories, we recruited four people to serve as feedback providers for the experiment. We chose a sample size of four to minimize variability in the feedback that would be returned to the participants while allowing the providers to experience and compare each composition condition. The providers were recruited from Upwork because of its large user base of experienced freelance writers and infrastructure conducive to coordinating extended tasks. All providers lived in the USA, had attained or were actively pursuing at least a Bachelor’s degree in creative writing or related fields, and had significant prior experience producing critiques of creative writing (e.g., as part of a creative writing studio or developing movie scripts). Providers were selected to encompass a range of professional and academic experiences , as shown in Table 5.2. Providers were remunerated \$155 for approximately 10-14 hours of story evaluations apiece.

Provider	P1	P2	P3	P4
Age	31	37	22	19
Gender	M	M	M	F
Highest Degree Pursued/Attained	BA	Ph.D.	BA	BA
Professional Experience	2 years	5+ years	1 year	<1 year
Writing Background	Fiction Critic	Literature Professor	Freelance Editor	Honors English Major

Table 5.2: Backgrounds of the feedback providers. All providers lived in the United States, were pursuing or had attained a Bachelor’s degree in creative writing or related fields, and had significant experience producing critiques of creative writing. Providers were instructed to compose feedback suitable for novice writers.

5.3.3 Writing Task

A creative writing task was chosen because it exemplifies the class of tasks for which a design solution is revised based on external feedback, it can be performed by a wide audience without requiring specialized software, and it promotes a sense of ownership of the content. The writing task consisted of *writing* and *revision* phases. In the writing phase, participants spent 10-15 minutes composing a 125-250 word story based on a writing prompt. During pilot testing, we determined this word range allowed for sufficient creative expression while enabling participants to finish within the time frame. Prior to drafting the story, half the participants were informed they would later receive feedback from a different participant in the study. The other half were informed they would receive feedback from an experienced writer. In the revision phase, each participant revised their story based on the feedback they received. The presentation of the feedback was based on the condition to which they were

assigned. The word limit was increased to 300 words to allow more room for revisions.

The task was administered through a self-contained web page created for the experiment. In both phases, the page consisted of the writing prompt, the task instructions, and a text field for composing the story. For the revision phase, the page also included any feedback the participants received, revision instructions, and a pre-filled form containing the participant’s initial story. Figure 5.2 depicts initial and revised stories from a participant in the rubric+scaffolding feedback condition with perceived expert source.

The task design was informed by pilot studies conducted on Mechanical Turk. A total of 48 participants took part in the pilot studies; those who participated in the pilot studies were not allowed to participate in the experiment. Through the pilot studies, we were able to improve the task instructions, usability of the task interface, and data collection.

5.3.4 Measures

To answer our first research question, we measured the revision depth between participants’ initial and revised stories, the effort participants put into revising their stories, the quality of their initial stories, and the quality of revisions to their revised stories. Revision depth was measured using word-level Levenshtein distance, which is less sensitive to small typographical edits than character-level Levenshtein distance. Revision effort was measured as the number of minutes participants spent revising their story, from making their first edit to submitting their revisions. A proxy for initial story quality was calculated by mapping providers’ rubric marks to scores (e.g., “Exceptional” = 4, “Above Average” = 3, etc.) for each criterion and summing these to obtain a final quality score. Revision quality was measured as improvement between initial and revised stories using randomized A-B comparisons, as elaborated upon in the Experimental Procedure.

To answer our second research question, we administered a post-task survey to participants after the revision phase of the study. The survey included items asking participants about the time and effort they put into their stories, the fairness and helpfulness of the feedback they received, and the perceived credibility and investment of their feedback providers (see Table 5.3 for examples). These questions were presented as 7-point Likert items (e.g., “The feedback I received helped me make effective, meaningful improvements to my initial story”; 1=strongly disagree, 7=strongly agree).

To answer our third research question, we recorded the amount of time feedback providers spent evaluating each story. This time was measured as the number of minutes between opening each story’s feedback page to submitting the feedback. Each provider was also administered a questionnaire upon completing their assigned evaluations. For each interface,

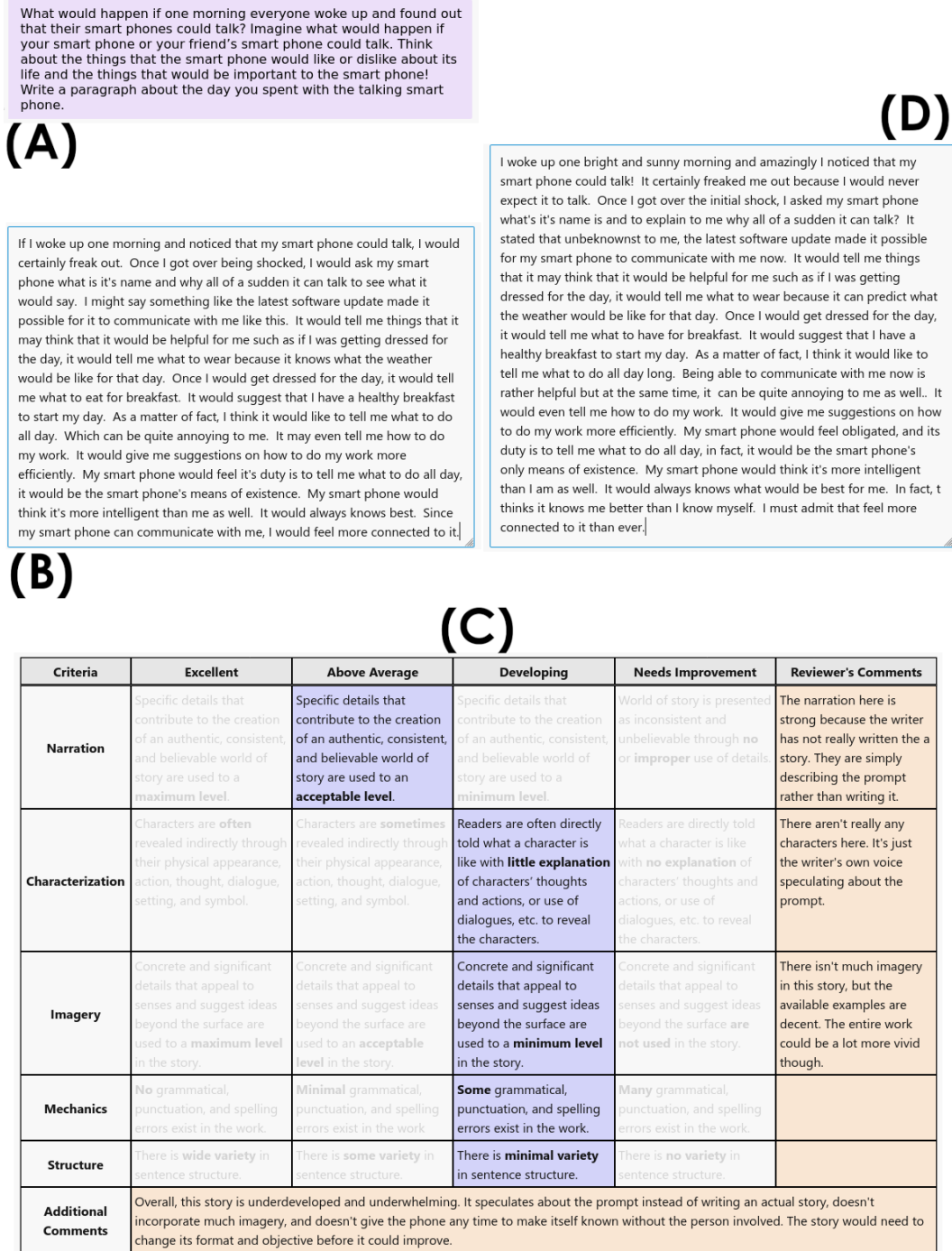


Figure 5.2: Overview of the experimental design. Participants reviewed a writing prompt (A) and composed a short story based on the prompt (B). Providers composed feedback for the participants (C), and participants revised their stories in response to this feedback (D). The workflow depicted is from the perspective of a participant in the rubric+scaffolding presentation condition. The rubric+open comments interface did not contain the rightmost “Reviewer’s Comments” column, and the rubric-only interface did not contain the “Reviewer’s Comments” or “Additional Comments” boxes. Interfaces were otherwise identical.

the questionnaire inquired about their processes for composing feedback, the time and effort they spent composing feedback, and how helpful they believed their feedback would be to recipients. Providers described their composition processes using open-ended text boxes. The remaining questions were presented as 7-point Likert items (1=strongly disagree, 7=strongly agree).

5.3.5 Experimental Procedure

In the writing phase, participants went through an informed consent process, completed the background survey, and created an identifier on the platform used for the experiment. They were then directed to the task web page where they read the task instructions and drafted their initial story based on the writing prompt. To track edit history, participants composed their story using the research platform and were instructed not to paste the content from a different tool or to reuse an old story of their own or anyone else’s. Participants were informed that upon submission, their story would be reviewed by either “a reviewer experienced in creative writing” or “another participant in the study”, depending on the source factor of their condition. Participants were also informed they would have a chance to revise and resubmit their story after receiving feedback, which would be provided within 48-72 hours.

In the evaluation phase, feedback providers were directed to an index of their assigned stories after reviewing consent information. Providers were briefed on the experimental procedure and the appropriate level of participant expertise to assume when composing feedback. They then reviewed the participants’ writing prompt and instructions for using each feedback composition interface. Providers were asked to evaluate participants’ stories in

Measure	Prompt
Credibility	“The person who wrote my feedback is knowledgeable about creative writing.”
Helpfulness	“The feedback I received helped me make effective, meaningful improvements to my initial story.”
Fairness	“The feedback I received was a fair assessment of the initial story I wrote.”
Investment	“The person who wrote my feedback wanted to help me improve my initial story.”

Table 5.3: Likert prompts corresponding to measures of participants’ feedback perceptions. Prompts were answered using a 7-point scale ranging from “(1) Strongly Disagree” to “(7) Strongly Agree”.

the order they appeared in the index, working at their own pace within the 48-hour evaluation period. After evaluating all of their assigned stories and completing the feedback composition questionnaire, providers were remunerated within 24 hours.

In the revision phase, participants were invited to return to the study platform and revise their stories in response to the feedback. Participants were presented with feedback corresponding to their assigned conditions, and revised their stories with respect to this feedback. After submitting their revised stories, participants completed the post-task questionnaire, and were remunerated within 48 hours.

We took precautions to minimize variance between feedback providers. All providers were instructed on how to using the grading rubric. Providers were further instructed to write 1-2 sentences in each of the per-criterion comment boxes explaining their rubric selections, and 2-3 sentences in the open comments box with any additional feedback for participants to improve their stories. Finally, providers also rotated through each of the interfaces in a different order to counterbalance ordering effects of interface on feedback composition.

To assess revision quality, we recruited two experienced writers from Upwork to review the initial and revised stories and perform A-B comparisons. We recruited a new set of reviewers due to concerns of fatigue and priming among the providers who evaluated the initial stories. We used A-B comparisons to assess quality differences between pairs of initial and revised stories. Similar to the feedback providers, reviewers were briefed and given instructions for comparing stories based on how well they followed the writing prompt. Initial and revised stories were presented side-by-side with random placement on each review page. Reviewers were asked to assess the relative quality of the stories using a 9-point Likert scale and a brief explanation of their rating (e.g., “stronger character writing, more polished dialogue”). Suggestions for using the scale were available as tooltips over each rating level, e.g.: “the right draft is better in small but appreciable ways that subtly improve the quality of the story.” Ratings were standardized for analyses such that negative ratings corresponded to stronger initial stories, while positive ratings corresponded to stronger revised stories. Reviewers were remunerated \$120 for 8-12 hours of evaluation each.

5.4 RESULTS

A total of 381 participants submitted a story in the writing phase of the study. Data from 46 participants was discarded for reasons such as plagiarism, off-topic work, or incomplete survey responses. Additionally, 50 participants did not return for the revision phase. This left 285 participants (144 female) who completed the study.

Initial story scores were normally distributed with a mean of 12.1 on a scale of 5 to 20,

	Rubric Only	Open Only	Rubric+Open	Rubric+Scaff.	F / χ^2
Init. Quality (RQ1)	11.6 (2.65)	12.2 (3.18)	12.1 (2.92)	12.6 (3.12)	1.804
Rev. Effort (RQ1)	19.8 (33.2)	13.8 (12.8)	16.4 (17.4)	17.6 (30.9)	0.549
Rev. Depth (RQ1)	102 (61.1)	101 (62.0)	101 (73.8)	98.5 (54.7)	0.067
Rev. Quality (RQ1)	1.36 (1.60)	1.50 (1.61)	1.71 (1.43)	1.92 (1.51)	9.501*
Credibility (RQ2)	5.19 (1.43)	5.33 (1.52)	5.84 (1.30)	5.89 (1.30)	17.40***
Helpfulness (RQ2)	5.15 (1.68)	5.56 (1.53)	5.71 (1.35)	5.82 (1.43)	9.475*
Fairness (RQ2)	5.62 (1.51)	5.62 (1.66)	5.80 (1.54)	6.05 (1.40)	5.589
Investment (RQ2)	5.48 (1.37)	5.90 (1.43)	6.14 (1.04)	6.03 (1.27)	14.55**
* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$					

Table 5.4: Mean and (SD) for measurements of participants’ perceptions and revisions, split by presentation condition. ANOVAs were used to analyze initial story quality (on a 5-20 point scale), revision effort (in minutes), and revision depth (in word edit distance), while Kruskal-Wallis tests were used to analyze the Likert data (all 7-point scales). F and χ^2 values shown are for main effects of feedback presentation; perceived source is omitted due to lack of significant main or interaction effects. Boldface indicates significant results. In general, perceptions of helpfulness, credibility, investment, and revision quality increased with level of feedback detail.

with a standard deviation of 2.95. Revision quality ratings were normally distributed with a mean of 1.62 on a scale of -4 to +4, with a standard deviation of 1.56. Stories spanned the full allowed range of 125-300 words, with a mean length of 262 words and a median length of 289 words.

In our analyses, we found no effects of perceived feedback source on any of our measures. For this reason, we do not include perceived source in our presentation of the data. We used the statistical software package R to perform all analyses. Likert data was analyzed using Kruskal-Wallis tests and post-hoc Dunn tests with Benjamini-Hochberg corrections. All remaining data was analyzed using ANOVAs and post-hoc Tukey HSD tests.

After adjusting for multiple comparisons, ANOVAs and diagnostic plots revealed the amount of time providers spent writing feedback and the perceived quality of that feedback was not significantly different among the four providers, nor did it change significantly over time. This suggests that the quality of feedback among our feedback providers was similar enough to proceed with our remaining analyses without further adjustment.

5.4.1 Revision Effort, Depth, and Quality (RQ1)

Table 5.4 summarizes the data for participants’ revisions and perceptions. A Kruskal-Wallis test revealed a significant effect of feedback presentation on story revision quality ($\chi^2(3) = 9.501, p = 0.023$). Revision quality in the rubric+scaffolding condition ($\mu = 1.92, \sigma = 1.51$)

was significantly higher than in the rubric-only condition ($\mu = 1.36, \sigma = 1.60, p = 0.015$). Nearly 48% of participants in the rubric+scaffolding condition received a quality rating of 3 or higher on their revised stories, indicating “substantial improvement.” By contrast, only 26% of participants in the rubric-only condition received a rating of 3 or higher. Our data shows a pattern of increasing revision quality as participants receive feedback at higher levels of detail. We will see this pattern reoccur throughout our analyses of participants’ feedback perceptions.

Two-way ANOVAs did not reveal any significant effects of feedback presentation on revision effort or depth. This suggests that while the quality of participants’ revisions increased with feedback level of detail, the quantity of their revisions did not.

5.4.2 Recipient Perceptions of Feedback (RQ2)

A Kruskal-Wallis test revealed a significant effect of feedback presentation on perceived credibility ($\chi^2(3) = 17.40, p < 0.001$). Credibility in the rubric+scaffolding condition ($\mu = 5.89, \sigma = 1.30$) was higher than in the rubric-only condition ($\mu = 5.19, \sigma = 1.43, p = 0.002$) and in the open comments-only condition ($\mu = 5.33, \sigma = 1.52, p = 0.037$). Credibility in the rubric+open comments condition ($\mu = 5.84, \sigma = 1.30$) was also higher than in the rubric-only condition ($p = 0.011$). The increase in perceived credibility with feedback level of detail reflects the earlier pattern seen with revision quality.

A Kruskal-Wallis test revealed a significant effect of feedback presentation on perceived helpfulness ($\chi^2(3) = 9.475, p = 0.024$). Helpfulness in the rubric+scaffolding condition ($\mu = 5.82, \sigma = 1.43$) was significantly higher than in the rubric-only condition ($\mu = 5.15, \sigma = 1.68, p = 0.021$). As with revision quality and perceived credibility, perceived helpfulness increased with feedback detail.

A Kruskal-Wallis test revealed a significant effect of feedback presentation on perceived investment ($\chi^2(3) = 14.55, p = 0.002$). Investment was significantly higher in both the rubric+open comments ($\mu = 6.14, \sigma = 1.04, p = 0.010$) and rubric+scaffolding ($\mu = 6.03, \sigma = 1.27, p = 0.007$) conditions than in the rubric-only condition ($\mu = 5.48, \sigma = 1.37$). Unlike other measures, perceived investment was highest in the rubric+open comments condition rather than in the rubric+scaffolding condition.

Perceived fairness was comparable between all four feedback presentation conditions. A Kruskal-Wallis test did not reveal any significant effects of feedback presentation on perceived fairness.

Likert Prompt	Rubric-only				Rubric+Open				Rubric+Scaff.			
	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4
“Composing feedback took a significant amount of time”	1	2	1	2	4	4	3	3	4	5	6	6
“Composing feedback took a significant amount of effort”	1	2	3	2	5	4	3	3	7	5	5	6
“I was able to effectively communicate my thoughts, insights, and critiques when composing feedback”	1	2	7	3	7	5	7	6	7	5	7	3
“I believe the feedback I composed will help the recipients make strong revisions to their stories”	1	2	7	3	5	6	7	6	7	6	7	5

Table 5.5: Providers’ responses to Likert prompts for questions regarding each of the three feedback interfaces. Prompts were answered using a 7-point scale ranging from “(1) Strongly Disagree” to “(7) Strongly Agree”. Providers agreed rubrics took the least time and effort to use, though they were the least effective for communicating their thoughts. Providers were more divided on whether the extra complexity of the scaffolding interface helped them write better feedback.

5.4.3 Value and Effort of Feedback Composition (RQ3)

In answering our first two research questions, we saw a pattern of increasingly higher quality revisions and favorable feedback perceptions by recipients in response to increasingly detailed feedback. We now contrast these findings with insights from the perspectives of their feedback providers.

Table 5.5 summarizes providers’ responses to Likert prompts. Of the three interfaces, the rubric-only UI required the least effort, but was perceived as the least effective by the providers. While providers self-reported composition took considerably longer in the rubric+open comments UI, most providers agreed it was more expressive and more helpful to recipients than the rubric-only UI. The rubric+scaffolding UI took more self-reported time and effort than both other interfaces. However, most providers did not believe they were able to write feedback any better than with the rubric+open comments UI.

A one-way ANOVA revealed a significant effect of feedback interface on actual evaluation time ($F(2, 282) = 28.52, p < 0.001$), reinforcing providers’ self-reported times. Compared to the rubric-only UI ($\mu = 2.99min., \sigma = 2.15min.$), evaluations took 31% longer in the rubric+open comments UI ($\mu = 3.92min., \sigma = 2.53min., p < 0.001$), and over twice as long in the rubric+scaffolding UI ($\mu = 6.06min., \sigma = 3.66min., p < 0.001$). These results reflect providers’ survey responses.

When asked whether their processes for using the grading rubric changed when working with the three different interfaces, providers reported a variety of differences and challenges:

It was easier for me to use the rubric when there was only one box versus when I was trying to find a way to fill four separate ones. – **P4**

...when I couldn't comment I took the rubric suggestions more directly. – **P1**

...I think I became modestly more generous with the judging rubric over time...in part because of the ambiguity of determining what exactly constitutes something like story world or imagery, and an erring towards generosity when more studiously considering these. – **P2**

Responses suggest the primary caveat of using rubrics stems from the forced selection among limited pre-authored statements, none of which necessarily reflects the provider's actual opinion. This disconnect may make selecting between rubric options more ambiguous and subjective than desired, in turn making it difficult to write comments justifying these selections.

When asked whether their processes for writing comments changed between the rubric+open comments and rubric+scaffolding UIs, P4 and P2 cited the extra step of breaking their thoughts down by rubric criterion, while P3 mentioned writing more concrete comments:

I created a series of questions to ask myself when I was working with the [rubric+scaffolding UI], but when I was working [with the rubric+open comments UI], it was easier for me to formulate my response on the rubric itself. – **P4**

The [rubric+scaffolding] interface directed my comments towards more specifically considering the three additional areas (narration, characterization, imagery), whereas the [rubric+open comments] interface tended towards me making holistic comments. – **P2**

The more detail oriented interfaces required me to answer in more concrete detail. – **P3**

Prior work has found holistic, high-level comments (P2) are more valuable at earlier stages of revision while concrete suggestions (P3) are more valuable at later stages [22, 150, 151, 152]. As a result, concrete per-criterion feedback may be more helpful at later stages of revision than at earlier stages.

When asked if any of the interfaces were particularly problematic to work with, providers were skeptical the rubric-only UI would be useful to writers, and reported both benefits and drawbacks to using the rubric+scaffolding UI:

I think the [rubric+scaffolding UI] was both best and trickiest...[it] pushed me into balancing somewhat more specific feedback with my tendency toward more totalizing feedback, and was the most thorough of the interfaces. The [rubric-only UI], I think, would not be particularly helpful in a creative writing course. – **P2**

I didn't really like the [rubric+scaffolding UI] because I believe having to write out explanations for each sections just takes away from what should really be said in the "overall comments." I think having just the one text box is the way to go. Easier to do and easier to get the ideas all together. – **P4**

5.5 DISCUSSION

The primary goal of our study was to explore the relation between the costs of composing feedback at various levels of detail and the corresponding benefits of the feedback to its recipients. Based on our findings, we offer an emergent framework for selecting appropriate feedback composition and presentation techniques depending on the needs of feedback providers and recipients (*italicized text denotes attributes of interest*).

[Provider] Feedback Composition Costs: We recommend using rubrics alone if one wants to minimize the cost of feedback composition (e.g., to scale feedback generation). Recipients perceived rubrics to be just as fair as the other feedback formats, and helped most recipients make improvements to their work comparable to those who received open comments.

[Provider] Perceived Helpfulness: We recommend using rubrics with open comments to maximize a provider's satisfaction with the feedback they compose. Providers found the rubric+open comments UI better than the other UIs for communicating their insights and composing feedback they believed would be helpful to recipients. Providers also felt writing comments was easiest and most natural in the rubric+open comments UI.

[Recipient] Revision Quality: To maximize revision quality, we recommend using a rubric with comments on each criterion. Compared to the other tested feedback types, feedback composed with the rubric+scaffolding UI led to the highest quality ratings for revisions. This type of feedback was also perceived as the most helpful by the recipients.

[Recipient] Revision Depth and Effort: If one wants to prioritize revision effort or depth, our recommendation is to use rubrics. Recipients made revisions of comparable depth

and effort regardless of their assigned feedback condition. However, the shorter time required for providers to compose feedback with rubrics alone compared to the other tested UIs may allow for quicker, more frequent feedback-revision cycles.

[Recipient] Satisfaction with Feedback: Our recommendation for eliciting the most favorable feedback perceptions from recipients is to use rubrics with comments on each criterion. Of all the conditions tested, recipients rated this type of feedback as the most fair, credible, and helpful.

[Recipient] Broad, Holistic Feedback: Whenever holistic feedback is desirable, we recommend using open comments. Providers specifically mentioned the rubric+open comments UI was more conducive to writing holistic comments than the other UIs. Because holistic feedback is especially valuable during early stages of design when revisions are large and frequent [22, 150, 151], open comments may be more appropriate and cost-effective than other types of feedback at these stages.

[Recipient] Specific, Concrete Suggestions: For recipients seeking concrete suggestions for improvement, we recommend using rubrics with comments on each criterion. The structure of our rubric+scaffolding UI guided feedback providers towards writing specific, concrete comments for helping recipients improve their work. The value of this feedback may be especially salient at later stages of revision when these detailed suggestions are most applicable [22, 152].

Designers of feedback exchange platforms and instructors can utilize these recommendations to tailor feedback to best suit the needs of their community or course. For example, in an online platform, the system could adaptively select the feedback composition interface based on attributes of interest selected by providers and recipients. This adaptive approach may be especially useful for creative activities integrated into platforms as a form of tangential play, e.g. to improve morale [153]. In these contexts, small design choices such as including ratings can stifle creativity and reduce satisfaction with creative activities [154], making choice of feedback interface an even more important factor. Adaptive implementations could also enable further analysis of which feedback attributes providers and recipients find most valuable, and how these differing values evolve into patterns of use. The recommendations in our framework should be considered preliminary because they were derived from a single study (see Limitations).

Compared to writing open comments, scaffolding feedback composition with per-criterion comments increased the reported time and effort providers in our experiment spent composing comments. While some providers believed open comments were more helpful to recipients than per-criterion comments, providers did not believe per-criterion scaffolding hindered their expressiveness, nor did recipients find per-criterion comments less personalized or invested

in their work. By contrast, prior work suggests scaffolding may reduce the difficulty of composing feedback at the expense of decreased expressiveness and personalization [142, 143, 144, 145]. This contrast may indicate the type of scaffolding used for composing feedback may affect composition and revision outcomes more than the presence of scaffolding alone. We recommend future researchers and designers of feedback exchange platforms compare and contrast how alternative forms of scaffolding affect feedback composition and utilization.

Our experiment involved experts providing feedback to a primarily novice population. Novices have been shown to write feedback almost as good as experts when given rubrics, but may not be as proficient at writing open comments [13]. Consequently, feedback composed by novices (i.e., peers) may have lead to different results in our experiment. While we believe the constraints on our writing task and control over perceived feedback source allows some degree of generalization, subsequent research should investigate how the production and utilization of authentic peer feedback varies with different feedback composition interfaces.

5.6 LIMITATIONS AND FUTURE WORK

We did not find significant effects of perceived feedback source on participants' feedback perceptions or subsequent revisions. This result might be attributed to the fact that our study did not differentiate the content of the feedback based on its source. We manipulated only the perception of the source to control for differences in content that might occur due to differences in expertise. A future study might explore how feedback that differs in scope, perspective, and quality affects recipient perceptions and revision outcomes.

Our study explored how different compositions of feedback affected recipients' perceptions of that feedback and revisions for a short writing task. Future work is needed to test the generalizability of our findings for tasks of longer duration, with multiple revision cycles, and in additional domains such as graphic design and programming. Future work is also needed to expand our framework to include additional measures of interest for feedback exchange, such as the learning that occurs from receiving and composing different forms of feedback.

We solicited participants from Mechanical Turk, where users typically participate in research studies for financial gain. This incentive structure differs from online feedback exchange platforms where users are driven by enjoyment, or classroom environments where learners are driven by grades. Future work is needed to test if participants driven by other motives such as learning or task enjoyment yield similar results.

Beyond future work addressing these limitations, we see several avenues for extending our research. Additional experiments could explore resolving situations where the attributes of interest selected by content creators and feedback providers yield conflicting recommendations

(e.g., they each point to the use of a different composition interface). Another direction for future work is to explore how to optimize for combinations of attributes simultaneously. For example, we found that extending rubrics with open comments cost only one additional minute for composition as opposed to three additional minutes when adding per-criterion comments to rubrics. The resulting feedback was perceived as comparably helpful to per-criterion comments by both providers and recipients, and resulted in revisions of comparable quality.

5.7 CONTRIBUTIONS

Feedback can be composed with many techniques that lead to different levels of detail and personalization of the content. In an online experiment, we compared provider perceptions, recipient perceptions, and recipient revisions between four different presentations of feedback. Recipients of rubrics with per-criterion comments perceived their feedback to be the most helpful and performed revisions that received the highest quality ratings. By contrast, providers believed per-criterion scaffolding added unnecessary complexity to composing feedback, suggesting open comments would be more helpful to recipients. Finally, rubrics alone took the least time for providers to compose, and still helped recipients make improvements to their work. We distilled these results into an emergent framework for selecting feedback presentations based on composition and revision outcomes of interest, such as using a rubric with per-criterion comments to maximize revision quality. We hope our results enable instructors and designers of feedback exchange platforms to make decisions that best balance the distinct needs of feedback providers and recipients.

A paper reporting the results from the study discussed in this section has been accepted for publication to *Computer-Supported Cooperative Work 2022*. This work makes three main contributions. First, we contribute empirical knowledge of how the presentation of feedback influences a recipient’s perceptions of the feedback and subsequent revision outcomes. Second, we offer insights linking these perceptions and outcomes with the feedback’s composition cost and perceived value from the provider’s perspective. Finally, we contribute an emergent framework for selecting feedback composition techniques based on specific attributes of interest and insights relating the costs of composition to the benefits of the resulting feedback.

The findings from this study revealed that while creators’ perceived feedback helpfulness was correlated with feedback detail, the differences in helpfulness were small enough that other factors (such as the time and effort required for providers to compose feedback) still warranted careful consideration. While my prior two studies examined how user interfaces influence a creator’s perception, interpretation, and usage of a single piece of feedback,

additional factors may be important when two or more pieces of feedback are involved. Particularly, the organization of multiple pieces of feedback and the mechanisms by which a creator navigates that feedback may influence the extent to which creators are able to interpret and act upon that feedback. This observation motivated my next study of how presenting feedback through an interactive visualization can help creators interpret, reconcile, and operationalize large collections of feedback.

Chapter 6: Exploring How Visualizing Feedback’s Topic and Opinion Structure Influences Feedback Exchange

In the preceding section, I investigated how presenting feedback at different levels of detail influences the feedback’s composition, the provider’s and recipient’s perceptions of the feedback, and the revisions performed in response to the feedback. My findings suggested that feedback may have value at any level of detail depending on the needs and constraints of feedback providers and recipients. While this experiment examined how participants revised based on a single piece of feedback at a fixed level of detail, in practice creators often receive feedback from multiple providers and at varying levels of detail. Large collections of such feedback can be especially challenging for creators to interpret due to the presence of several potentially conflicting perspectives from different feedback providers.

To better assess how user interfaces can help creators interpret large collections of feedback, in this section I present a field study of how students in a UI design course leverage an interactive visualization tool to explore and implement themes and suggestions from large collections of feedback. This study synthesizes tool interaction logs, data from three surveys, and interviews with students (N=12) and teaching assistants (N=2) to explore which tool features and feedback metadata learners find most valuable in processing feedback from multiple providers on a semester-long course project. By incorporating this interactive visualization into the instruction of an authentic user interface design course, I investigate how and to what extent presenting feedback through a visualization enhances and extends student’ strategies for interpreting, engaging, and utilizing the feedback they receive from their instructors and classmates.

6.1 INTRODUCTION

Design-based learning is a powerful pedagogical approach that prepares students to develop solutions to real-world problems [155]. This learning style typically involves an iterative process where students create and refine prototypes based on feedback from stakeholders [1, 41, 135]. The feedback collected from different stakeholders such as peers, instructors, and external audiences is difficult to interpret because it often contains varying topics, opinions, and structure [156, 157, 158]. This challenge is exacerbated in learning contexts because students often skim formative feedback [45], lack concrete strategies for interpreting feedback [33, 34], and lack the opportunity or volition to implement those strategies [34]. Additionally, this challenge persists even when the individual pieces of feedback are of high quality [9, 16, 32].

Two threads of prior work most relevant to this study address the challenge of feedback interpretation. One thread explores using cognitive interventions such as reflection [16, 47] or paraphrasing [46] to make sense of feedback. These activities require individual, subjective effort that does not readily scale to large feedback collections or to multiple recipients (e.g., teammates). A second thread explores aiding feedback interpretation through tool support by presenting interactive visual summaries to help users extract key themes from the content [7, 9, 159, 160]. However, existing tools typically assume feedback is generated with a known structure (e.g., using rubrics or fixed response options), limiting their utility for discovering patterns in unstructured formative feedback common in design-based learning.

The present work intersects and extends these threads by deploying a new type of feedback visualization tool in a project-based user interface design course and studying how students used the tool to make sense of peer and instructor feedback. The tool structures free-form feedback by visualizing how it maps across feedback providers, topics, and opinion types (e.g., praise or suggestions). The tool associates each feedback statement with an icon and arranges these icons in a grid indicating each statement’s author (row), topic (column), and opinion (icon shape and color). We designed a field study to determine 1) what interpretation goals students pursue when reviewing formative feedback, 2) what patterns of tool use emerge in pursuit of those goals, 3) how the tool affects students’ feedback review processes, and 4) how using students’ own labels to generate the visualization compares to using a third party’s labels. Though the tool was studied in a prior controlled experiment [22], that work did not study the tool’s use in the context of users’ own projects and feedback, patterns of tool interaction, or impact of labeling on users’ review processes and visualizations.

In the course, 18 teams of 3-5 students each developed a 12-week user interface design project of their choice. Our study targeted three critical graded deliverables that would likely most benefit from revision in response to feedback: the project proposal (weeks 1-2), low-fidelity prototype (weeks 6-7), and functional prototype (weeks 10-11). Students presented an initial version of each deliverable in an online studio session while peers and a teaching assistant wrote free-form feedback. Teams were given one week to review their feedback using the tool and revise each deliverable in response to the feedback. To study the tradeoffs of different labeling approaches, the research team labeled each student team’s feedback for the project proposal. Students could revise the research team’s labels for the low-fidelity prototype, and labeled the feedback themselves for the functional prototype. After each revision period, we surveyed students to learn about their goals for interpreting feedback, how they used the tool to accomplish these goals, and which aspects of the tool they found most valuable. At the end of the course, we interviewed 12 students about how reviewing feedback with the tool compared to their typical review processes, how these processes changed when

using the tool, and how they felt about the different labeling approaches. To complement the students' perspectives, we interviewed the course's teaching assistants to learn how using the tool affected their instruction, grading decisions, and perceptions of their own feedback.

Our investigation revealed that student teams used the tool to find the feedback that was valuable to them, assess the quality of their project deliverables (i.e., critical problems and key successes), prioritize and discuss which feedback to address, and justify their design decisions. These goals were primarily accomplished by exploring a multitude of patterns of opinion icons in the visualization and reviewing the providers' backgrounds. Students also used the icon patterns to order their review of feedback details (e.g., begin with critical statements and end with praise). Students perceived that the benefit of the visualization outweighed the cost of labeling the feedback, and expressed that labeling the topics and opinions themselves prompted them to critically analyze each feedback statement. The distribution of labels assigned by the students was also consistent with the distribution of the research team's labels, suggesting comparable reliability between the two approaches. Several students reported adapting strategies they learned using the tool (such as color coding feedback by topic or highlighting repeated statements) when reviewing feedback for projects external to the course. The teaching assistants reported leveraging patterns in the visualization to aid grading decisions and offer teams additional project guidance during office hours. The teaching assistants also expressed considering how their feedback would be presented in the tool when writing it, such as not being overly critical to avoid an imbalance of red icons and not dissenting too far from majority opinions to avoid concerns over grading decisions.

This work makes three contributions to the HCI community. First, we show interactive visualization techniques can be applied to help students discover, prioritize, and share critical issues in the feedback for their creative projects. The techniques implemented in the tool and resulting design implications from this study should generalize to interpreting unstructured feedback received in myriad contexts such as in online critique communities, academic peer review, and job performance evaluations. Second, we dispel concerns surrounding the costs of having students provide the meta-data needed to generate a feedback visualization. Students found that the benefit of generating the visualization in the tool outweighed the cost of labeling the content (at most 60 minutes per feedback collection). Students also indicated a preference for labeling their own feedback because they found that labeling aided their comprehension of the feedback and they could appropriate the labels to match their own organizational style. Third, the collective data generated from students' tool interaction has the potential to aid instruction, e.g., by allowing instructors to determine whether students are writing feedback with an appropriate balance of opinions and topics and to

curate examples of feedback that promote actions by the recipient. This work advances the idea that feedback *interpretation* is as important as feedback composition, and shows how interactive visualization tools can be leveraged to aid interpretation tasks and create new opportunities to teach specific feedback interpretation strategies.

6.2 RESEARCH QUESTIONS

We deployed the visualization tool in a course and had students in the course use it to interpret feedback for their course project. The field study was designed to answer four research questions:

- **RQ1:** What goals do students try to accomplish when using the feedback visualization tool and what patterns of use emerge to accomplish these goals?
- **RQ2:** What processes for reviewing feedback do students retain from using the tool, and what difficulties arise when reviewing feedback without the tool?
- **RQ3:** How does having students label their own feedback impact the resulting visualizations and their perceptions of the tool?
- **RQ4:** How can a feedback visualization tool be best appropriated for instructional purposes?

Answers to these questions will contribute to a base of knowledge and best practices for designing technology aimed at helping end users interpret a collection of feedback written by multiple providers.

6.3 FEEDBACK VISUALIZATION TOOL STUDIED

We extended the design and implementation of a Web-based feedback visualization tool (Decipher [22]) for the classroom study. The tool adds structure to a collection of unstructured written feedback by visualizing how the feedback maps across providers, topics, and opinions. The goal of the tool is to help users discover useful patterns in the feedback that would be difficult to extract from the text alone – such as patterns of praise and criticism across topics or between providers – and to develop an action plan based on these patterns to improve the project.

The main page in the tool displays the text of the collection of feedback for the project on the left and presents the visualization of that collection of feedback on the right (see

Figure 6.1). In the visualization, the feedback is organized in a grid by reviewer (rows) and by topic (columns). Topics are sorted by their frequency in the feedback, with the most frequent topic placed in the leftmost column. For each topic referenced by a provider, a graphical icon indicates if the piece of feedback consists of praise (thumbs up, green), criticism (thumbs down, red), a suggestion (bulb, yellow), a question (question mark, yellow), a neutral comment (filled, yellow), or a mixed opinion (half praise, half criticism). The cell is empty if a provider does not reference that topic in their feedback. Icon colors and shapes were used in tandem to improve visual accessibility.

The text and visualization are linked through interaction. Selecting a piece of feedback (e.g., a sentence) on the left highlights the graphical icon associated with that piece of feedback in the visualization. Likewise, placing the cursor over an icon in the visualization causes the tool to scroll to and highlight the corresponding feedback on the left. Clicking on an icon displays a window containing the associated piece of feedback and interaction for choosing an intended action (*must do*, *discuss it*, *disagree with it*, and *consider it*) for the piece of feedback and a checkbox for marking the intended action as completed. The user can search for keywords in the feedback and for pieces of feedback labeled with specific intentions. Search results are displayed by dimming the icons that do not match the search criteria.

The tool imports the collection of feedback and attributes of the providers from a CSV file. Once imported, the tool partitions the feedback into sentences. A user can then open the annotation page in the tool to label the topic and opinion of each sentence, and can recursively merge two or more adjacent sentences if desired. This meta-data is necessary for the tool to generate the visualization shown in Figure 6.1. Default topics are defined in the tool and, in our study, they were derived from the rubric associated with the corresponding project deliverable. For example, topics for the project proposal included project goals, solution alternatives, user audience, and feasibility. A user can add new topics in the tool, whereas the set of opinion categories was fixed and could not be revised by the user in the current implementation.

The attributes of a provider are displayed in a tooltip when the user places the cursor over the icon representing that provider in the visualization. For our study, the attributes were: familiarity with the project topic from Very Unfamiliar (1) to Very Familiar (5), major (computer science, social science, humanities, etc.), and year of study (junior, senior, grad student, or teaching assistant). The user can sort the visualization by these attributes (e.g., to find the feedback from the teaching assistant or from those most familiar with the project topic). A short document describing the visualization and interaction features of the tool was linked from a Help button placed on the main page of the tool.

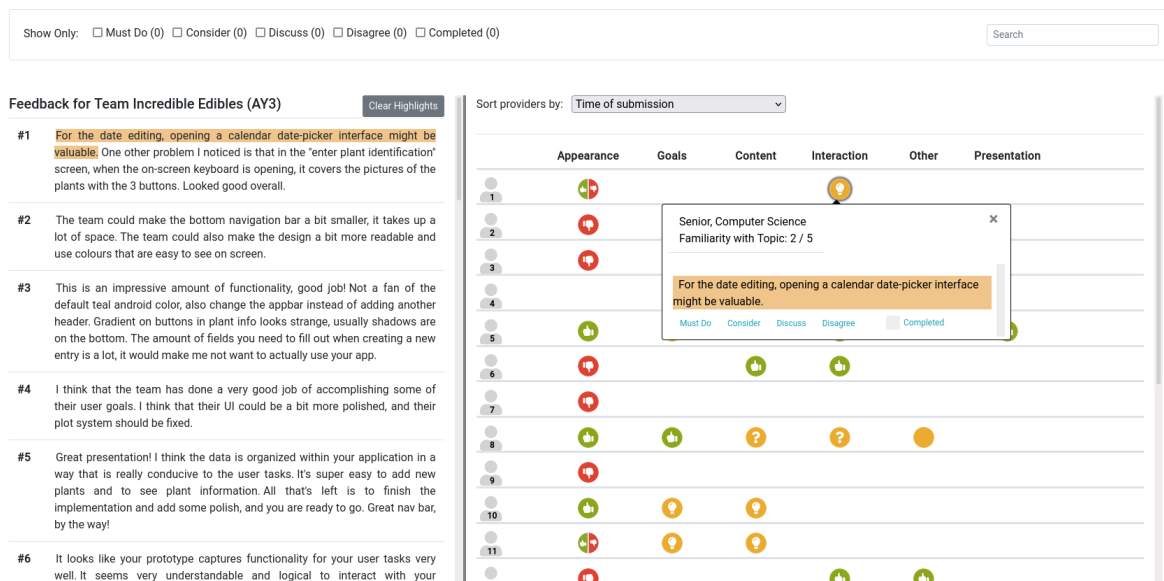


Figure 6.1: The main user interface screen of the feedback visualization tool, currently showing a collection of feedback (left) and its visualization (right) for a team’s functional prototype. The graphical icons on the right depict the opinions in the feedback by provider (the rows) and by topic (the columns). Hovering the cursor over an icon highlights the corresponding piece of feedback on the left, and vice versa. Clicking an icon displays a tooltip showing the feedback provider’s background information, the piece of feedback on the left corresponding to the topic, and the intention labels for tracking feedback status. The top of the main user interface screen includes a row of checkboxes for filtering feedback by intention labels, a search bar to filter feedback by keywords, and an option to sort the feedback by attributes such as a provider’s self-reported familiarity with the project topic, year of study, and major.

6.4 METHODOLOGY

To answer the research questions, we deployed the feedback visualization tool (Section 6.3) in a project-based user interface design course taught at a public university in the U.S. A field study method was chosen so that we could study the use of the tool in authentic feedback review processes. The course was chosen because of its emphasis on having students gather and act on feedback as part of the iterative design process, and the specific tool was chosen because it visualizes the type of unstructured written feedback students already exchange as part of their project work in the course. In the course, students completed user interface design projects in teams and used the tool to review feedback written by classroom peers and the teaching assistants for three project deliverables, each with an initial and revised version. Multiple uses allowed students to gain familiarity with the tool and how to best incorporate it into their feedback review processes. Using the tool for three project deliverables was the

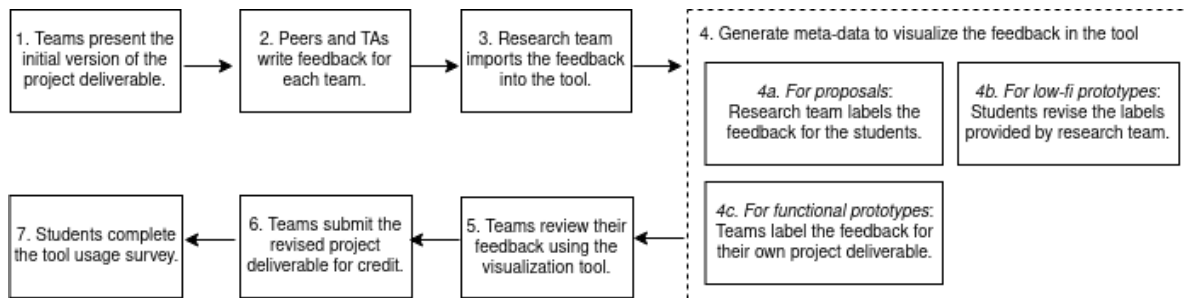


Figure 6.2: An overview of the study procedure. The steps shown in the diagram were repeated three times, once for the project proposal, low-fidelity prototype, and functional prototype. Student teams presented the initial version of a project deliverable during their studio section. Peers and the TA wrote feedback for each team in an online form. The research team imported all the feedback into the tool. For the proposal (project weeks 1-2), the research team labeled the topics and opinions in the feedback and student teams could access the resulting visualization in the tool without additional work. For the low-fidelity prototype (project weeks 6-7), student teams could revise the labels assigned to the feedback by the research team to best fit their own interpretation of the content. For the functional prototype (project weeks 10-11), students labeled the feedback on their own and a default set of topics were available in the tool. Student teams then reviewed the feedback in the tool, submitted a revised deliverable for course credit, and completed a tool usage survey. After the project ended, members of the research team interviewed students and the two teaching assistants in the course.

most that could be practically integrated into the course.

The research questions were addressed by analyzing data collected through surveys, interviews, and interaction logs. In addition, for the third research question, the research team progressively transitioned the labeling of the topics and opinions in the feedback from the research team (first use of the tool) to the students (final use of the tool). This approach helped students understand why these labels were needed by interacting with the visualization first, and allowed the students to compare their experience with the tool when having to label their own feedback vs. having the feedback labeled by an external party.

An author of this study was the instructor of the course. The instructor and teaching assistants were blind to which students consented to allow their data to be used for the purpose of research. Members of the research team not affiliated with the course performed the interviews and collected all consent information. The study was approved by the Institutional Review Board at our university.

6.4.1 Design Course and Projects

The user interface design course targets upper-level undergraduate and beginning graduate students studying computer science. There were a total of 98 students (34 female, 64 male) in the course. About 10% of the students were from other departments on campus such as chemistry, health, psychology, and the arts. For most of the students, this was their first course on user interface design. The lecture topics included user research, prototyping, implementation, and evaluation techniques. The instructor organized students into teams of 4-6 students to balance skill sets. There were a total of 18 teams in the course. The teams applied the lecture topics to a 12-week user interface design project of their choice. Examples of these projects include a mobile app and hardware to remotely start a motorcycle, an app for scheduling the first and subsequent vaccine shots in one scheduling session, and an app for homeowners to request assistance with shoveling the snow from their sidewalks and for volunteers to respond to those requests.

The projects were structured as a design process. There were a total of nine project deliverables, with one deliverable submitted each week for a grade. Teams presented the deliverables during a weekly design studio. Each studio section had 20-30 students, making up 5-6 teams. Teams were assigned using the CATME [161] software to balance team compositions based on students' skill sets. Presentations were performed via video conferencing technology due to the pandemic. Teams presented their project deliverables in the studio while classmates and the teaching assistant wrote feedback in an online form. The teaching assistant also facilitated five minutes of oral critique before proceeding to the presentation of the next project in that studio section.

Three project deliverables created by the student teams were targeted in this study: 1) the project proposal (project weeks 1-2), 2) the low-fidelity prototype (project weeks 6-7), and 3) the functional prototype (project weeks 10-11). The proposal was a document describing a project idea, including the user need, existing solutions, and user audience. The low-fidelity prototype consisted of sketches or mockups representing a first-cut solution for the project. The functional prototype was a programming implementation of a team's low-fidelity prototype. These three deliverables were selected for the study because they represented key milestones in the project timeline and incorporating a revision cycle for these deliverables would benefit teams the most. These deliverables were each split into an initial and revised submission. Teams presented the initial submission in studio, received a web link to access their feedback in the feedback visualization tool, revised the submission based on the feedback, and submitted the revision for course credit the following week. Teams could only access the feedback for the initial submissions in the tool. The revised submissions

earned three times more credit than the initial submissions to incentivize attending to the feedback and making changes. Additional project deliverables such as user research reports were not included in the study design because the course timeline would not accommodate an initial and revised submission for each of the other deliverables. Teams did gather additional feedback from potential end users for other project deliverables. It was not practical for students to gather end user feedback in time to visualize it alongside the peer and TA feedback for the deliverables that were included in the study.

6.4.2 Students

At the onset of the course, students completed a survey about their experience receiving feedback for open-ended work (N=76). In the survey, students described a recent project (not the project in the current course) which was revised based on the feedback received from multiple people, the format of the feedback, and the providers who wrote that feedback. Of the respondents, 89% reported having received feedback from multiple people for open-ended work two or more times while 11% reported no prior experience. Examples of open-ended work for which the students received feedback included written reports, programming projects, and senior design projects. For the instances described, 81% of the students reported that the collection of feedback was from 2-5 providers, most often classroom peers (74%) and teaching assistants (14%). In the instances described, the format of the feedback was written (22%), verbal (25%), or a combination (53%). Students wrote that what made the collection of feedback difficult to understand was 1) knowing which statements in the feedback to prioritize for revising their work (e.g., to receive the highest grade), 2) translating vague and ambiguous statements in the feedback to actual revisions of their work, 3) remembering all the feedback received (particularly if received verbally) and what it meant in relation to their work, and 4) forgetting to address some feedback because it was received through multiple channels.

6.4.3 Tool Usage Surveys and Interaction Logs

Students completed three tool usage surveys, once after using the tool to review the feedback associated with the initial submission for each of the three project deliverables included in this study. Students completed the surveys individually, in part to allow each student to decide whether to give consent for their data to be used for the purpose of research. Students earned course credit for completing the surveys.

The tool usage survey associated with the project proposal had 20 questions related to our

study. A student estimated the time spent reviewing the feedback in the tool and described how the team reviewed the feedback in the tool (e.g., as a group, or individually and then as a group). The student then responded to 14 Likert scale questions about the usefulness of the the tool and each of its main user interface features (exploring topics, exploring opinions, seeing provider backgrounds, labeling intended actions, etc.), the quality of the topic and opinion labels provided by the research team, the quality of the feedback, and how difficult the feedback was to understand. The questions were phrased as a statement (e.g., “My team found the feedback visualization tool useful for understanding the feedback received.”) and the responses were on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree). In 4 open-ended questions, the student described the most useful insights the team discovered in the feedback, how the features of the tool were leveraged to find these insights, and the most significant benefits and problems the team experienced when using the visualization tool for reviewing the feedback.

The surveys corresponding to the usage of the tool for the low-fidelity and functional prototypes asked the same questions as described above. However, for the low-fidelity prototype, teams could revise the topic and opinion labels that were provided by the research team. The survey for this deliverable included two additional Likert questions about revising the labels (“My team revised the initial topic and opinion labels that were provided to us in the tool”) and the importance of doing so (“My team found it important to revise the initial topic and opinion labels to best reflect our own understanding of the feedback statements”).

For the functional prototype, the research team did not label the feedback. Each student team was instructed to use the tool to label the topics and opinions in the feedback they received on their own. The extent and appropriation of the labeling was up to them. This survey also asked about the perceived trade-off between the effort required to perform the labeling and the benefits of being able to visually explore the feedback on a scale from 1 (cost outweighs benefit) to 7 (benefit outweighs cost) and asked students about their preference for using a feedback visualization tool on a similar project in the future on a scale from 1 (prefer a traditional text viewing tool) to 7 (prefer the feedback visualization tool). Finally, the survey asked students to describe how the use of the feedback tool affected the process by which they review the feedback, compared to how they might have reviewed the same feedback if in text-only form.

The feedback visualization tool was instrumented to log students’ interactions with the tool. This included the topic and opinion labels assigned to the feedback, the interactions with the graphical icons, the intention labels assigned, and the searches performed on the feedback.

6.4.4 Student and TA Interviews

After the project was complete, the research team conducted semi-structured interviews with 12 students and the two teaching assistants (TAs) in the course. Student interviewees were recruited through an open call placed in the final tool survey and sent to the course mailing list. Student interviewees consisted of 4 juniors, 5 seniors, and 3 graduate students. Seven were Computer Science or Computer Engineering majors, three were minoring in CS, and two were Health Technology majors. None of the students had previously used a similar feedback visualization tool. The TAs were Ph.D. students in Computer Science in the area of human-computer interaction. The interviews were conducted over Zoom, lasted 30-60 minutes, and were screen recorded with consent. Interviewees were each remunerated \$40.

To begin the student interviews, we wanted to observe their process for reviewing feedback when that feedback was presented in a text viewing tool (Google Doc) and compare it to the processes observed and described when reviewing similar feedback in the visualization tool. We adapted a poster design and a collection of feedback for that poster that was developed in a prior study [22]. The feedback was written by users recruited from a micro-task platform. There were a total of six pieces of feedback for the poster and the word count was comparable to the word count of the feedback students typically received for a project deliverable in the course. We also confirmed the feedback collection varied in topics and opinions but was also scoped such that it could be reviewed during the first 15 minutes of the interview.

We asked the student to imagine helping the project’s creator discover the most useful insights in the feedback and to show us their process for reviewing the feedback in a document editor (Google Docs), using any of the features in the editor desired. They were also asked about the strengths and weaknesses of reviewing the feedback in a document editor and to describe how they might have used the visualization tool to review this same collection of feedback.

For the next part of the interview, the student accessed the feedback for their project deliverables in the visualization tool. The student located what they recalled to be the most interesting insights in the feedback from any of the deliverables, and to show us how they discovered these insights with the tool. In addition, we asked about their first impressions of the tool, perceptions of labeling the feedback in the tool, and how the use of the tool helped them develop new processes (if any) for reviewing a collection of feedback written by multiple people.

For the TA interviews, we inquired about how the feedback they wrote was typically returned to students in other courses, their impressions of the tool, how the use of the tool helped (or hindered) their ability to grade, write feedback for, or mentor the projects, how

#	Content		Topic	Opinion
5-0	The prototype definitely does address your project goals.		Select from existing topics Goals ▾ Add a topic <input type="text"/>	praise ▾
5-1	The only thing about the project may be feasibility.	Merge Above	Select from existing topics Goals ▾ Add a topic <input type="text"/>	criticism ▾
5-2	It seems that there are quite a few features associated with each task.	Merge Above	Select from existing topics Interaction ▾ Add a topic <input type="text"/>	criticism ▾
5-3	Maybe instead of listing palettes in the accessibility tool try to find something a bit more visually appealing.	Merge Above	Select from existing topics Appearance ▾ Add a topic <input type="text"/>	suggestion ▾
5-4	Overall great job on the prototype.	Merge Above	Select from existing topics Presentation ▾ Add a topic <input type="text"/>	praise ▾

Figure 6.3: One page of the feedback labeling interface as seen by a student team for the low-fidelity prototype deliverable. Each page displays the feedback written by a single feedback provider and splits it into individual feedback statements. Students are able to assign an opinion and topic to each statement from predefined lists, define custom topics for the feedback statements, and merge related feedback statements together.

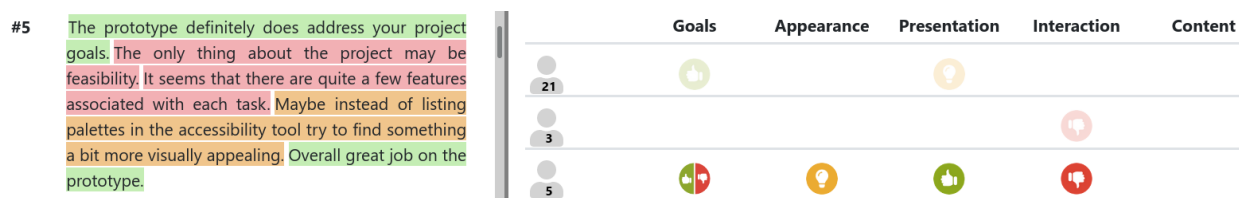


Figure 6.4: How the labels assigned to the feedback in Fig. 6.3 appear in the visualization's feedback review interface. Icons are arranged in a grid according to each feedback statement's assigned topic (column), feedback provider (row), and opinion (shape and color). Hovering over a feedback statement on the left highlights the associated icon on the right, while clicking the statement retains this highlighting, allowing students to highlight and compare multiple pieces of feedback at the same time.

they felt about their feedback being visualized alongside the student feedback in the tool, and how they believed that this might have affected the feedback they wrote.

6.4.5 Procedure

The instructor informed students that a feedback visualization tool would be used in the course and demonstrated the tool prior to its first use. When teams presented their initial project proposals in the studio, the other students and the TA in that studio section wrote feedback and answered the background questions (e.g., familiarity with the project topic) in an online form. The form instructed students to write one paragraph of detailed feedback for the project deliverable, including its strengths and weaknesses, and to explain their reasoning. The form also suggested topics to cover in the feedback based on the rubric associated with that project deliverable.

At the conclusion of the studios that week, the research team aggregated all the feedback into a CSV file and imported it into the feedback tool. The research team and teaching assistants then labeled the topics and opinions in the feedback. The topics were derived from the rubric used for grading the deliverable, while the set of opinion labels was fixed (see Section 6.3). The research team and teaching assistants labeled the feedback together for one project deliverable for calibration, then divided the labeling of the feedback for the remaining projects. Each student team received a Web link to their project’s feedback in the tool within two days of the completion of the recent studio. A student team received feedback from 20-25 peers and the TA. The student team reviewed the feedback in the tool, revised the project deliverable based on the feedback, and submitted the revision the next week for course credit. Students then completed a tool usage survey.

The above process was repeated for the low-fidelity and functional prototypes, but the strategies for labeling the feedback were adjusted to evaluate the tradeoffs between having an external party label the feedback and having the students label the feedback on their own. For the low-fidelity prototype, student teams were encouraged to revise the labels assigned by the research team to best fit their own interpretations of the content or needs. For the functional prototype, students were instructed to label the topic and opinions in the feedback on their own. A default list of topics was provided in the tool based on the rubric for the deliverable, but the teams could add new topics to this list. After this last use of the tool, the research team interviewed students and the teaching assistants about their usage and perceptions of the tool.

After the completion of the interviews, the research team followed the generalized inductive approach [162] to develop and iteratively refine a coding schema. The interview and open-

Label	# Students	Description	Example Idea Unit
Similarities	47	Focusing on insights regarding topics of interest repeated by multiple providers (e.g., by looking at clusters of icons)	“We attempted to look for patterns in feedback from our peers to ensure that we didn’t skew our revisions based on one person’s response.”
Ordering	46	Structuring feedback review in a certain order (e.g., criticism, then suggestions, then praise)	“We clicked on the arrows and first looked at red arrows which pointed what was wrong followed by the green good arrows for what is working on [our] prototype.”
Expertise	34	Sorting or otherwise locating feedback based on a peer’s seniority or topic expertise (including finding TA feedback)	“Based on the seniority of the student, we could see that younger students were not able to offer as much feedback on feasibility but had opinions on design instead.”
Proportions	41	Using proportions of opinion icons to identify assess a project’s strengths and weaknesses	“The more useful aspects of the visualization tool are the culmination of comments and the icons under each category because we could quickly gauge how well our prototype was [received].”
Strengths	5	Reviewing praise to identify strong areas to maintain in the revision	“Seeing the green thumbs up icon when you first open it up makes you feel good since it indicates that you are on the right track you just need some tweaks.”
Discuss	26	Insights from the tool sparked or reinforced discussions among team members	“We looked at all the critical feedback (using the right-hand side of the tool mostly) as a team and discussed/made a list of the changes we would make as we went.”
Debate	4	Features of the tool (e.g., assigning intention labels) prompted team to debate relative importance of feedback	“I found the “must do/consider” buttons to be most useful. It forced us to debate the feedback we received.”
Familiarize	17	Tool helped learner familiarize themselves with the feedback they received	“...we really got to understand how our peers viewed our demo and which suggestions or critiques were the most important to discuss and work on.”
Match	60	Tool’s features matched the learner’s goals (e.g., tool was intuitive or helpful)	“Having the labels of “Must Do”, “Consider”, “Discuss”, etc were helpful for understanding the tasks that were important and needed to be completed compared to extra or unnecessary tasks.”
Mismatch	33	Tool’s features mismatched the learner’s goals (e.g., tool overwhelmed the learner with information or features)	“labeling the feedback sometimes seemed like a lot of extra effort compared to the benefits of visualizing the feedback.”
Process	9	Tool helped learner develop a new process for reviewing or engaging with future feedback	“Since we can label fragments of each feedback paragraph, we can label different parts with different topics, and use it as a way of dividing work between team members.”
Change	43	Learner wanted something added to, removed from, or changed about the tool (feature requests, QoL fixes, bug reports, etc.)	“If you could implement a “comment counter” type of system and then sort the comments based on which comments match with the most commented stuff, that would be even more helpful.”

Table 6.1: The coding schema applied to the interview and open-ended survey responses. Each idea unit in both the survey and interview data was assigned exactly one of the labels in the schema. Only the labels in the second column were directly used for the coding.

ended survey responses were partitioned into 1083 idea units. Two members of the research team coded the idea units according to the schema. To test inter-rater reliability, a sample of the data (80 units) was coded by both raters. The remaining units were split between the two raters and coded independently. Table 6.1 summarizes the schema applied to the interview and open-ended survey responses.

6.5 RESULTS

Eighteen teams received feedback for each of three project deliverables. For each deliverable, a project team typically received 20-25 pieces of peer feedback and one piece of feedback from the TA. On average, one team's collection of feedback contained 1447 words (SD=397) and was split into 76 units of feedback (SD=19) categorized into 5-9 topics (see Figure 6.1 for an example). Students rated the feedback they received from their peers helpful (mean=5.9, SD=0.9, 7-pt scale), and felt they were able to determine how best to revise their project deliverables using this feedback (mean=5.6, SD=1.0, 7-pt scale). Students reported using the tool for 20-60 minutes on average for reviewing the feedback for each project deliverable. About half the students reported using the tool individually to review feedback before meeting with their team to discuss it, while the other half reported reviewing the feedback in the tool as a team. This was accomplished through the use of screen sharing in a video conferencing tool, since all teamwork was conducted remotely. Table 6.2 shows a summary of responses to the tool usage survey for the final deliverable. Responses to the other two tool usage surveys have similar results and are omitted for brevity.

A total of 69 students (42 male, 27 female) consented to their data being used for the purpose of research. We answer our research questions by drawing from the student surveys and interviews, the data collected through the tool, and the TA interviews. In the next subsections, we use the notation S# to refer to student survey respondents, I# to refer to student interview participants, and TA# to refer to TA interview participants.

6.5.1 Goals When Using The Tool (RQ1)

We found from the interviews and open-ended survey questions that students aimed to accomplish three goals when reviewing a collection of feedback with the tool: find the valuable feedback, assess project quality, and facilitate communication and coordination amongst teammates.

Question	Mean	SD
1. The feedback from peers was helpful for improving the project deliverable this past week (1=strongly disagree, 7=strongly agree).	5.76	0.94
2. The feedback from the TA was helpful for improving the project deliverable this past week (1=strongly disagree, 7=strongly agree).	6.37	0.76
3. My team found understanding the collection of feedback overwhelming (1=strongly disagree, 7=strongly agree).	3.18	1.43
4. My team found the feedback visualization tool helpful for understanding the feedback received for the project deliverable this past week (1=strongly disagree, 7=strongly agree).	5.27	1.12
5. My team was able to discover useful insights in the feedback when using the tool (1=strongly disagree, 7=strongly agree).	5.64	0.87
6. My team discussed the insights discovered in the feedback when using the tool (1=strongly disagree, 7=strongly agree).	5.77	0.98
7. My team was able to find the specific issues in the feedback that we were most interested in when using the tool (1=strongly disagree, 7=strongly agree).	5.51	1.08
8. My team found the feedback visualization tool helpful for determining how to best revise the project deliverable based on the feedback received (1=strongly disagree, 7=strongly agree).	5.47	1.03
9. My team found it useful to assign the topic and opinion labels to the feedback statements (1=strongly disagree, 7=strongly agree).	4.64	1.50
10. My team would like to use the same tool to review feedback on a future project deliverable (1=strongly disagree, 7=strongly agree).	5.18	1.33
11. Please rate the usefulness of exploring the topic columns in the feedback (1=not useful, 7=very useful).	4.99	1.28
12. Please rate the usefulness of exploring the opinion icons in the feedback (1=not useful, 7=very useful).	5.04	1.28
13. Please rate the usefulness of reviewing the background of the feedback providers (1=not useful, 7=very useful).	3.60	1.69
14. Please rate the usefulness of labeling the intended actions (e.g., “must do”) for each statement in the feedback (1=not useful, 7=very useful).	4.08	1.58
15. Please rate the usefulness of marking the intended actions as Complete (1=not useful, 7=very useful).	3.86	1.60
16. Please rate the usefulness of labeling the topics and opinions in the feedback (1=not useful, 7=very useful).	4.90	1.30
17. Please rate the usefulness of reviewing the feedback in order on the left side of the screen (1=not useful, 7=very useful).	5.45	1.20
18. If you needed to review feedback for a similar project in the future, please rate your preference for using a text viewing tool (e.g., a Google Doc or Web page with the feedback text) or a feedback visualization tool like the one used in the course (1=prefer text tool, 7=prefer feedback visualization tool).	5.37	1.31
19. Please rate how you feel about the trade off between the effort required to perform the labeling and the benefits of being able to visually explore the feedback (1=cost outweighs benefit, 7=benefit outweighs cost).	4.69	1.35
20. Estimate the total degree of change that you believe the team made from the initial to the revised project deliverable by addressing issues in the feedback (1=no change, 7=significant change).	5.25	1.15

Table 6.2: The mean and standard deviation for students’ responses to the rating questions from the third tool usage survey. The data from the first two tool usage surveys showed similar ratings and is omitted for brevity. Responses were structured as 7-point Likert items, with endpoints shown in each question. Q11-17 were presented as a grid in the actual survey.

Finding the Valuable Feedback

Students found the tool helpful for finding statements in the feedback that they perceived to be valuable for their project deliverable (Q5: mean=5.64, SD=0.87; Q7: mean=5.51, SD=1.08 in Table 6.2). Based on responses to the tool usage surveys and interviews, we identified three main patterns of use that students reported for finding the valuable feedback. One pattern reported by 46 students (67%) in the surveys and 5 students (42%) in the interviews was structuring their review of the feedback in a particular order based on the types of icons in the visualization. The typical order was to review the critical statements first, then the suggestions and questions, and end with the statements of praise. Students stated that using the critical statements as the initial entry point into the feedback collection helped them identify problematic areas of their projects, while reviewing suggestions afterwards directed them towards potential solutions.

We first check all the red labels, then classify them, and extract the ones we think it is interesting and reasonable. Then check the yellow label, classify and discuss, and finally go through other checks to see if there is any missing information. [S38 (Team 16), Lo-Fi Prototype Survey]

We always go to the red thumbs down, because like those are the negative things...the criticisms we can deal with, so we always, always go find where the red thumbs down are, and we try to take a look, and then we just see what it is saying and discuss it over our team call. and after that, we go to the insightful [suggestion] button here...we consider this neutral like, it's not really bad...they don't really say anything bad about the feature, they just say "oh maybe you can do this maybe you can do that"...we just put that into consideration...other than that we just skimmed through it. [I2 (Team 10)]

A second pattern for finding the valuable feedback was observing vertical and horizontal clusters of the same category of opinion icons in the visualization. These clusters indicated redundant critiques of a topic or a reviewer who was particularly critical or supportive. This pattern was mentioned by 47 students (68%) in the comments they wrote in the tool usage survey and by 3 students (25%) in the interviews. Students also gave positive ratings for exploring the feedback by topic (Q11: mean=4.99, SD=1.28) and by opinion (Q12: mean=5.04, SD=1.28) in the tool.

...we would first look out for the red thumbs down...the column or the row with the most, those were obviously the ones we looked at first...the columns with

the most feedback on it are usually the areas you need to pay some attention to, unless it's all thumbs up. So we definitely knew that...we had the backend and everything running, but we hadn't so much focused on the UI part of it...that's basically how we approached it...first look for all the negatives, and then also look for which column has the most amount of feedback given for it. [I1 (Team 18)]

When we met up as a group, the first thing we would do is look at the visualization side of it, and see which columns had the most notes. So appearance and interaction...seems like where we spend the most time, and appearance has a lot of critiques on that....we would then look at that and say "okay, let's make note of all of these things...let's read through them." [I6 (Team 5)]

A third pattern for finding valuable feedback reported by 34 students (49%) in the surveys and 4 students (33%) in the interviews was prioritizing feedback based on reviewers' self-reported familiarity with the project topic and year of study.

One of the interesting patterns we discovered is that people who are less familiar with the topic don't seem to care about existing solutions as much as those who are more familiar with it. We found out this by using the sorting feature of the tool. [S56 (Team 11), Project Proposal Survey]

We found interesting differences in the feedback from different levels of students. Based on the seniority of the student, we could see that younger students were not able to offer as much feedback on feasibility but had opinions on design instead. [S37 (Team 9), Project Proposal Survey]

Among those who prioritized feedback review by year of study, several students reported reviewing critical statements from the TA first, and then reviewing peer feedback on those same topics for further elaboration and to find suggestions for improvement. Other students explored these relationships in the feedback in the opposite order, using the feedback from the TA to confirm or elaborate issues mentioned in the peer feedback. In both cases, the organization of the feedback by topic and opinion in the tool further enabled these explorations.

We found that the feedback from the TA was a very strong benchmark and very whole at identifying a broad range of issues and concerns with our project, both good and bad. All the reviews from my peers touched upon at least one concern raised by the TA and expanded upon their viewpoint. These patterns

were apparent because of the identifying likes, dislikes, and neutral hand icons which helped us quickly sort through the reviews and understand the sentiment across the board for a particular issue based on the hand icon colors. And clicking on the hand icons helped us highlight the review which causes that hand color. This helped us a lot. [S9 (Team 10), Project Proposal Survey]

So I went through all the students first and the TA was the last one. and it either confirmed something...if all the students were saying this one thing and then the TA agreed with them, it almost put more weight on that...versus if a couple students said something and the TA was like “actually I think this is good because x y z”...I think this is good then that’d be a point of discussion with my group mates. [I6 (Team 5)]

In sum, students found the valuable feedback by reviewing criticisms and suggestions before other icon types, observing criticisms that were repeated by multiple providers, and prioritizing feedback from peers who reported high familiarity with the project topic and from the TAs. However, exploring the reviewer background in the tool was rated as neutral (Q13: mean=3.60, SD=1.69), possibly indicating only some students used the reviewer background as a way to find the feedback that was valuable to them. The usage patterns described in this section may not be exhaustive, as students may not have recalled all the different uses of the tool and new patterns might develop with additional experience with the tool. Also, individual students likely performed different subsets of these patterns based on the feedback content, the project deliverable, and their own preferences.

Using clusters of icons for assessing project quality

Forty-one students (59%) in the surveys and 5 students (42%) in the interviews reported using the visualization to evaluate the quality of specific aspects of their project deliverable or its overall quality. Many students mentioned observing the proportion of praise, criticism, questions, and suggestions to evaluate the quality of their project deliverable and estimate the amount of work that would be required for the revision. Students reported visually scanning the criticism icons in each topic column to quickly determine which areas of their projects were in most need of attention. Similarly, 5 students (7%) in the surveys and 1 student (8%) in the interviews reported observing clusters of praise icons for assessing which aspects of their project deliverable they should avoid changing while revising other aspects of their work.

[The tool] helped me quantize the amount of praise comments we had, and the amount of improvement comments we had. We were able to obtain a better summary gauge without reading every single feedback, which is something we couldn't have done on an all text feedback form. [S24 (Team 5), Functional Prototype Survey]

In terms of the praise, we would all just go through alone and look through general things. I think in our app, we had a good sense of people really liked our ui and obviously the app looks great, so what we knew was that any changes we make, we can make them, but let's not do it at the cost of changing our colors, our layout, the way it just looks when people enter the app, because clearly people really like that. [I5 (Team 16)]

Students often reported that they reviewed opinion icons to identify the strengths and weaknesses of their projects before reading the associated text. In this sense, the distribution of the icons themselves functioned as a high-level form of feedback summarizing the detailed text feedback.

Communicating and coordinating revisions

The third main goal that 26 students (38%) in the surveys and 5 students (42%) in the interviews reported pursuing with the tool was facilitating communication and coordination amongst team members. Students reported using the tool's feedback organization features to structure their team discussions, as well as using feedback trends to develop their arguments during these discussions. For example, when reviewing peer feedback with which they disagreed, students were prompted to justify their decisions not to address certain issues raised in the feedback.

After the annotations were all done, we would sort of discuss based on what feedback we got. So if there were any questions or suggestions on what we should do...during our group meeting...we would discuss the feedback and use that as a basis of any design changes we would make...we'd include our own opinions on top of that, like, if someone gave feedback and we all very strongly disagreed with that, then we might not follow it as much. But we mainly used it as a way to discuss how we could do changes. [I11 (Team 7)]

We tried to address all suggestions and criticism in the feedback, either by actually implementing changes, or by explaining the reasoning behind our design choices

for the things we did not change in our presentation. [S55 (Team 1), Functional Prototype Survey]

Four survey respondents and one interview participant stated that assigning intention labels to their feedback forced their teams to debate the relative importance of each piece of feedback, which helped them prioritize revisions more effectively and make sure they addressed the most important feedback.

I found the “must do/consider” buttons to be most useful. It forced us to debate the feedback we received. [S59 (Team 12), Project Proposal Survey]

...whenever we take a look at [our feedback]...we’d click on “disagree” because of the project scope and how we only have like two weeks to develop it, then we just have to disagree with this current feedback. So I really think that these “must do”, “consider”, and “discuss” [labels were] really helpful for our team, because we could just say “hey what should we do with this specific feedback?” [I2 (Team 10)]

Students leveraged the above mechanisms to communicate with their teammates and coordinate revisions for their project deliverables. Although the tool was designed primarily for individual use, the behaviors some students demonstrated in this study highlight several opportunities for the tool to facilitate team-based feedback review.

6.5.2 Comparing Feedback Review with and without the Tool (RQ2)

As part of the post-project interviews, we performed a short deprivation protocol to study which strategies (if any) performed with the tool continued when reviewing feedback in a text-only form. The interviewees used a document editor (Google Docs) to identify what they believed to be the most important insights in a collection of six pieces of feedback for a design project unrelated to their own course projects. Interviewees were encouraged to use any of the tool’s features they felt necessary to mark up the text or take notes. Interviewees were given 20 minutes to review the feedback, extract what they felt were the most important insights, and think aloud while reviewing the feedback.

We observed that most students performed strategies similar to those implemented in the visualization tool and created a visible trace of their processing of the feedback. For highlighting, two students applied color-coded highlighting to identify different topics in the feedback, and two other students used color to differentiate criticism from suggestions.

Five students used a single color to highlight important insights, while the remaining three students did not use color to highlight at all. Regarding additional strategies, two students copied the most critical statements from the feedback to the end of each piece of feedback, and in one case a student added topic labels (e.g., “Content:” or “Presentation:”) and notes to the copied statements. Similarly, a different student selected feedback statements and added comments in the format “[topic] [opinion]”. Three students used boldfacing to mark the feedback statements they found most valuable. Two students reviewed the feedback without marking up the feedback text in any way. While students mimicked some of the same techniques implemented in the visualization tool, they weren’t able to apply these techniques as effectively without the tool’s consistent visual structure. For example, students who used highlighting typically developed arbitrary mappings of the colors to the topics or opinions in the feedback, and did not consistently apply these mappings.

Following the deprivation protocol, interviewees were asked about the strengths and weaknesses of reviewing text-only feedback, as well as the strengths and weaknesses of the visualization tool. Interviewees cited the difficulty of remembering (N=7) and organizing (N=4) the feedback as the biggest weakness of reviewing feedback in text-only form. By contrast, interviewees cited the ability to aggregate and categorize data (N=5) and identify the most common and important criticisms (N=5) as the biggest strengths of the visualization tool. These responses reflect the value of the visualization tool in its facilitation of different strategies and visualization of topic and opinion structure in a consistent manner. Interviewees also cited the ability to develop their own interpretations of the feedback without possible biases imposed upon the feedback’s presentation as the biggest strength of reviewing feedback in a text viewing tool over the visualization tool (N=6).

When asked in the second part of the interviews whether their strategies for reviewing feedback changed after using the feedback visualization tool for the course projects, 6 interviewees reported developing new processes for engaging with feedback, many of which they felt would carry beyond the use of the tool. One student mentioned they began highlighting plain-text feedback they received with different colors to help them review and organize the feedback more effectively.

I really liked the idea of color coding. Usually whenever I go in and look at anyone’s feedback and it’s just in text form, I would probably like make notes on a separate document for it but I wouldn’t visually go in and change the way the feedback looks, but I think that helped me process things a lot easier when I went back into the feedback after the first time I looked at it. So I know that’s something I’ve kind of even implemented in general, like actually going

and working and annotating with the text instead of like on a separate document.
[I10 (Team 2)]

A different student reported that after working with the tool, they began taking notes on recurring themes in plain-text feedback they received to reduce the burden of reviewing and reprocessing it later. Another student who was in the habit of skimming plain-text feedback stated that after working with the tool, they learned to appreciate the value of finding and marking the commonalities in the feedback they received, noting they wanted to ascertain that they were making full use of the feedback.

When going through feedback I would honestly skim through multiple pieces of feedback and try to see if there were any common elements. If I did see common elements then I would kind of jot them down, and every time that I saw something similar to what I read before I would just kind of increase the value of that comment, just so I could remember that more people were arguing for this approach. That I feel like is a better way of going through feedback than just trying to remember it all in your head and trying to proceed from there because that just gets very confusing very fast [I1 (Team 18)]

It definitely changed a little bit. Before, like I mentioned I just go through a couple of them and then also, I probably only see big themes...like big commonalities between them. But I think now if I were just given like a bunch of text based feedback...I kind of go through line by line and then really think about what smaller commonalities each of these reviews have. And then focus on those a lot more and highlight them so I don't [forget] them, which is something I never used to do...and just make sure I'm fully utilizing everything that people have written for me. [I4 (Team 11)]

A fourth student who had previously considered praise feedback unhelpful reported that they more thoroughly considered the praise they received in the context of consciously maintaining stronger aspects of their design.

...a lot of time when people are given feedback, they're so worried about fixing the bad things...that sometimes they do it at the cost of changing something that was good in their app or just good in their work that they did. So I think something that I've really learned is that looking at good feedback and understanding what someone says about something that you did well is super important and can actually help your weaknesses and help you not make changes that you shouldn't be making. [I5 (Team 16)]

With the feedback visualization tool, students found the feedback that was valuable to them by observing clusters of the same category of opinion icon, among other strategies. Without the tool, we found that some interviewees (N=10) also attempted to process the topics and opinions in the feedback, but used ad-hoc color codes, inline comments, or free-form notes, while others (N=2) processed the feedback in memory. These results suggest that user strategies for processing feedback might have limited effectiveness without the support of a tool designed to foreground the topics and opinions in the feedback with a consistent visual presentation.

6.5.3 Impact of Labeling Approach on Visualization and Feedback Perceptions (RQ3)

The feedback tool requires meta-data in the form of topic and opinion labels in order to generate the visualization for a feedback collection. An important consideration is how this meta-data should be created and by whom. We compared three approaches: the research team labels the feedback, student teams revise the labels assigned by the research team, and students label the feedback on their own. We begin the results with the student-led labeling.

Student teams labeled their own feedback for the third project deliverable. There were a total of 1096 statements in the peer and TA feedback, or about 61 statements for each team. The interaction data showed that the student teams fully labeled all the statements in their respective collections of feedback. We also found that the distribution of labels assigned by the student teams was similar to the distribution of the labels assigned by the research team for the second project deliverable (see Table 6.3). Although different collections of feedback were labeled by the students and by the research team, the similar distributions suggest that students did not interpret the feedback for their own projects with a particular bias (e.g., with overly positive or negative interpretations).

On the tool usage survey associated with the third project deliverable, 60% of respondents

	Praise	Criticism	Neutral	Suggestion	Question
Proposal	62.33%	33.41%	4.26%	N/A	N/A
Low-fidelity Prototype	48.99%	17.75%	0.00%	26.52%	6.74%
Functional Prototype	50.80%	13.93%	0.00%	30.26%	5.01%

Table 6.3: Distribution of opinion categories by project deliverable. The research team labeled all the feedback for the first project deliverable (Proposal), teams could revise the labels assigned by the research team for the feedback for the second deliverable (low-fidelity prototype), and student teams labeled their own feedback for the third deliverable (functional prototype). The “Suggestion” and “Question” labels were not yet implemented in the tool at the time of the first project deliverable.

estimated their team spent at most 30 minutes labeling the feedback, and 30% estimated their team spent between 30-60 minutes. When asked how their team labeled the feedback, 58% of respondents indicated that one member of the team labeled the feedback, 25% indicated that the team labeled the data together, 10% indicated that the team divided the work, and 7% did not answer or reported other strategies. Students agreed with the statement that labeling the topics and opinions in the feedback was useful (Q16: mean=4.9, SD=1.3) and perceived the benefit of seeing the resulting visualization to outweigh the cost of labeling the feedback (Q19: mean=4.7, SD=1.4).

When asked about the tradeoffs between the three approaches experienced for labeling the feedback, interviewees (N=12) mentioned two benefits of labeling the feedback on their own. First, nine interviewees mentioned that labeling the feedback forced them to think more carefully about each piece of feedback they received and to become familiar with the feedback more quickly than in the prior project deliverables. The remaining three interviewees did not cite any particular advantages to labeling the feedback themselves over having the research team label it for them. In response to the open-ended question about the tool's main strengths, survey respondents (N=17, 25%) also noted the benefits of familiarizing themselves with the feedback through the labeling process. Students noted that assigning the labels prompted them to begin thinking about the most critical pieces of feedback before they saw the resulting visualization.

...for the third iteration where we were like in charge of labeling our own feedback that was like...we were like oh it actually you know it's like not tedious but like a process you kinda have to go through each sentence and really think about "oh this is for appearance, this is for content, I think this is positive, I think this is negative, this is an idea, this a confusion point"...I think that was like where we probably got the most value out of 'cause we were thinking about it a lot more and actually working with this a lot more and categorizing it for ourselves rather than just agreeing with what was already there. [I4 (Team 11)]

The biggest benefit was that we actually got very familiar with the feedback as we organized it. When it was organized for us, we didn't pay attention to every piece of feedback, but this time we really got to understand how our peers viewed our demo and which suggestions or critiques were the most important to discuss and work on. [S25 (Team 2), Functional Prototype Survey]

A second benefit was being able to appropriate the labels to fit their own needs. For example, two students reported in the tool usage survey for the third deliverable that their

respective teams developed a new schema for the topic labels in the tool, appropriating the topic labels for use as an action list.

The opportunity to categorize feedback to help us prioritize the issues was helpful. Also, we split up the team into frontend and backend using a divide and conquer approach. Therefore, the categories allowed us to easily delegate work. [S5 (Team 16), Functional Prototype Survey]

For disadvantages of labeling the feedback on their own, eight interviewees cited issues with the labeling process, including the difficulty deciding upon the appropriate label for a piece of feedback, difficulty tracking where they left off when splitting the labeling into multiple sessions, and the lack of an undo button. The other four interviewees cited the time required to label the feedback in the tool as the primary disadvantage. However, three of these interviewees also stated the value of the resulting visualization outweighed the labeling effort.

For the second project deliverable, students were given the option to revise the labels assigned by the research team. The interaction logs indicated that none of the project teams revised these labels. The lack of revision was likely due to the fact that students agreed with the statement on the tool usage survey that the labels assigned by the research team matched what they would have assigned themselves (mean=5.16, SD=1.06, from the second tool usage survey).

I don't think we changed any of the labels, to be honest. I think we did look through them. But we agreed with most of them. So we're like, "okay, cool." [I4 (Team 11)]

And even in the second time, where we had the ability to go back and edit, we didn't really use...that opportunity, because it was already kind of done for us. And we honestly got a little lazy, because it was already done. [I10 (Team 2)]

Despite not revising the research team's labels in practice, all interviewees preferred having the option to revise the pre-labeled feedback. When asked which of the three labeling approaches they preferred (pre-labeled without revision, pre-labeled with revision, and self-labeled), six interviewees stated they preferred labeling the feedback themselves, three preferred revising the research team's labels, and one felt both of these approaches were equally valuable. The remaining two interviewees had more nuanced opinions, stating pre-labeled feedback might be most effective for the first time they used the tool, or when they were working with a group they were less familiar with.

If you are comfortable with your team, if you work together often, I would say that doing it manually...is gonna be very helpful because you have the manpower to do it. and then you're enjoying your time anyways on the team call. it's not like a burden for you. but maybe some people who [don't] click with their teammates and just want to get over with everything quickly...if I were to put myself in their shoes, I would rather have the automatically annotated feedback just so I don't have to spend too much time with these random people...it really depends on a case by case basis. [I2 (Team 10)]

I would probably pick [self-annotation], but if you were just talking in general, I think it's a good idea to have [editable staff annotations] as an introductory approach. if it's not for me or for other students, it's nice to have an introduction to it and then have the ability to annotate, because I think you understand the tool a lot better that way [I10 (Team 2)]

In sum, from these different approaches for labeling the feedback, we found that students are willing to fully label the topics and opinions in the feedback for their projects in order to produce the resulting visualization, assign labels without inflating or deflating the criticism they receive, and report processing the feedback more carefully than if the labels are provided to them.

6.5.4 Appropriating the Tool for Instructional Purposes (RQ4)

During the interviews, the teaching assistants referenced several different uses of the tool for instructional purposes. TA1 mentioned using the tool to review student feedback to help synthesize her own feedback. She also reported writing her feedback precisely such that the topics in the feedback would be easy to recognize and label in the tool.

There were a couple of times somebody would say something...and I could kind of call that out in my feedback like "oh so and so mentioned I think this is a good idea" or I may say something like "oh I saw a lot of the feedback said something about this particular feature, maybe you want to consider doing something with that." So I sort of used it to inform the feedback I was writing. [TA1]

One thing is it sort of made me more aware of the categories... thinking more explicitly about the topics of the feedback and where it might fall on the visualization. So I might tweak the wording a little bit to make it more clear like "oh

this is about the idea and this is about the implementation”. So I guess it made me be a little bit more precise with my language. [TA1]

TA1 also mentioned that she used the tool in part to justify her grading decisions, stating she was more inclined to deduct points if a lot of criticism icons were present in the visualization.

I don’t think it hindered me at all. I think there may have been a couple times when we were using the tool where I sort of looked at the overall like color of the of the comments that were given here to see, is there a lot of green? I guess it helped with the severity of deducting points, if that makes sense. So I might be more inclined to deduct more points if I saw a lot of red and or yellow in the chart. [TA1]

TA2 mentioned revisiting the visual structure of her feedback in the tool (e.g., the critical points) to guide her later discussions with the project team during office hours. The same TA also reported being conscious of the need to balance the opinions in her critique of the project deliverable because she did not want her feedback to appear inconsistent with the peer feedback adjacent in the tool.

Whenever a student comes to my office hour [to] discuss about their project, I can’t remember all 10 teams’ feedback at that point sometimes, so I go to this tool, look at my previous feedback, and this tool has these nice visualizations that I can look into the negative feedback I gave to that team, so I just quickly look through what I said negatively about their project, and then remember what I said and give advice. [TA2]

...my impression of reading the comments from students was that they were mostly positive. They were very positive. compared to my feedback, because I have to have some negative points if I need to deduct points. I felt since everyone was so positive about the project, I felt kind of guilty to say negative things in my own feedback, so I tried to have a more positive tone and I also tried to make positive statements when I’m making feedback. [TA2]

These uses of the tool for instruction should be considered preliminary because our sample only included two teaching assistants. However, these preliminary findings do indicate an opportunity for future work exploring how a feedback visualization tool could aid course instruction.

6.6 DISCUSSION

We deployed a feedback visualization tool in a project-based course to learn how students leverage this type of tool for interpreting feedback received from multiple providers – in this case, classroom peers and the teaching assistants. Here we summarize the main findings for each of the research questions and discuss these findings in light of the tool’s interaction design, the learning context of the study, and the broader literature.

6.6.1 RQ1: Pursuing Feedback Goals

For RQ1, we found that students wanted to accomplish three goals when using the tool. One of these goals was finding the feedback that they believed was valuable for improving their project. This was accomplished by using the tool to explore feedback by its opinion structure (critical first), observing clusters of the same category of icon within the topic columns, and reviewing the background of the providers to weight their specific feedback statements. Students stated they typically read critical comments first to determine the problems they needed to address in their project deliverable, and sometimes ignored positive comments altogether.

The student’s preference to attend to the critical statements first conflicts with how providers are typically taught to write formative feedback. Providers are typically taught to write design feedback in the form of a “feedback sandwich”, offering praise before criticism and ending on a positive note [163]. The expectation is that the recipient will read the feedback in the order it was written. Prior work has found that reading the feedback from a positive to negative valence order improves the recipient’s perception of the feedback collection relative to placing the negative feedback in other positions [21]. Future work could experiment with different constraints to affect how recipients are able to review the feedback (e.g., requiring recipients to open all praise comments before critical comments or suggestions are revealed). The color scheme used for the opinion icons also made it easy for students to attend to the critical statements (red icons) before the praise and suggestions (green and yellow icons), as red hues are known to have a strong pop-out effect [164]. Future work could allow users to configure different color schemes for the icons. Presentational choices such as icon color and ordering constraints demonstrate the potential impact of separating feedback’s representation from its content, and encourage further exploration of how different means of presenting feedback can facilitate feedback review.

Students used their peers’ self-reported familiarity with the project topic and year of study to weigh feedback statements, but did not report using other attributes such as major of

study. The right amount of background information to collect and present in a feedback visualization tool is an open question. Additional background data such as gender or ethnicity could facilitate unwanted biases [165], whereas fewer details could inhibit the interpretation of the feedback. The most appropriate background information to present might differ by project topic and stage, which argues for giving instructors the option to toggle the collection and visibility of various background information for the feedback providers in the tool. The data set generated through the use of the tool (e.g., labels indicating how students intend to act on specific feedback statements) might also enable analysis that sheds light on how students respond to the feedback written by providers with different backgrounds.

A second goal students pursued with the tool was assessing the quality of their projects by comparing the relative proportions of different opinion icons across topics or within a particular subset of the topics. One way to further enhance the tool in support of this goal would be to implement visual summaries of the icon categories for each topic column and overall. Observing the ratios of icon categories in the tool might be difficult for large feedback collections, e.g., if using the tool to visualize the course evaluation comments written by hundreds of students. To scale to larger feedback sets without overwhelming the user with information, the tool could incorporate hierarchy into the visualization by organizing reviewers and topics into a taxonomy and allowing creators to “drill down” [166] to progressively explore the details. Finally, a third goal students pursued with the tool was facilitating team discussion and coordination. e.g., by assigning intention labels to the feedback statements and later searching and filtering the feedback by these labels. Team coordination could be further supported by allowing students to enter an estimated amount of time needed for making revisions prompted by the feedback statement, assigning the revisions to a team member, and generating an action list for each team member. The time estimates could be used to ensure balanced workloads between team members and to help prioritize the revisions in the project timeline.

6.6.2 RQ2: Processes for Reviewing Feedback

When reviewing text feedback during the interviews, students demonstrated techniques that mirrored the tools’ features, such as color coding praise and critical comments and marking intentions to act on statements by boldfacing those statements. To determine whether these techniques were indeed inspired by the tool, we performed the same interview protocol with five new students enrolled in a later instance of the same course which did not incorporate the visualization tool. These students were asked to think aloud while reviewing a given collection of feedback using a text viewing tool. While students in both sets of

interviews attempted to find the most valuable insights by looking for redundancy, students who did not use the visualization tool leveraged alternative techniques that required extra steps to process the feedback and identify these insights. Specifically, interviewees who did not use the visualization tool highlighted with arbitrary color mappings ($N=2/5$) or not at all ($N=3/5$), did not attempt to organize the feedback ($N=4/5$), and backtracked when reviewing feedback ($N=4/5$). By contrast, students who used the visualization tool typically highlighted the feedback as they were reading it ($N=10/12$), corroborating their reports of familiarizing themselves with feedback through the annotation process. These students also used color-coded opinion highlighting that mirrored how they looked for opinion clusters in the tool ($N=8/12$), and marked important insights with either boldface ($N=2/12$) or inline comments ($N=3/12$) similar to their use of the tool’s intention labels. Our findings suggest the visualization tool helped students develop general strategies for reviewing feedback which can be applied with or without tool support. These findings also highlight the opportunity to teach effective feedback review practices through tool design.

6.6.3 RQ3: Impact of Labeling Approach on Feedback Engagement

We found that students were willing to fully label the topics and opinions in their feedback to produce the visualization, assigned opinion labels without particular bias (such as preferring labeling feedback as praise or suggestions rather than as criticisms), and reported processing the feedback more carefully than if the labels were provided to them. Additionally, survey responses indicated that students believed the benefits of generating the visualization and familiarizing themselves with the feedback for their project outweighed the costs of assigning the labels. These findings support having students generate the meta-data for their own feedback when deploying similar feedback visualization tools in course contexts. However, course staff may want to initially label some feedback themselves to demonstrate the resulting visualization so students understand why the labels are necessary and can calibrate their labeling decisions. The course staff might choose to label all the feedback in situations where the instructors’ perspectives and consistency of labels between teams is most desirable.

The labeling interface in the tool provides a list of the feedback statements and pull-down menus for selecting the topic and opinion categories for each statement. The user interface is straightforward, but the statements are shown in the form of a list rather than in the form of a narrative. An alternative labeling interface might allow the user to label the data while reading the feedback content, similar to how the interviewees annotated the text in the deprivation protocol. For example, the tool could provide a specific highlighter for each opinion category and the user could select text and drag representations or copies of that

text into topic categories. The resulting visualization would be the same, but the labels could be captured in way that aligns with what was observed during the deprivation protocol in the interviews.

There are additional options for generating the meta-data that were not tested in this study. One option is to ask the providers to label the feedback as they write it, which could help them communicate their ideas and alleviate work for the recipients. However, this approach could also increase the provider’s cognitive burden when writing the feedback and lead to inconsistent labels across multiple providers. Another option is to use machine learning models to assign topic and opinion labels that the students could revise. We conducted an exploratory test of this approach with our data set. Using the students’ and research team’s labels as a baseline, our exploration found that an off-the-shelf model achieved a 59.4% accuracy on assigning opinion labels across all feedback collected in the study for each team and deliverable. This accuracy is likely a lower bound, as a training data set could be expanded over time as more feedback statements are labeled in the tool. Note that we did not have sufficient data to build and test statistical models for the topic labels because the topics were different for each project deliverable. A future study should compare how delegating the labeling task to different sources (provider, recipient, instructor, or automation) affects not only accuracy and effort, but also the recipient’s comprehension of the feedback.

6.6.4 RQ4: Instructional Impact of the Tool and Future Improvements

Both teaching assistants used each team’s visualization to help determine a grade for the associated project deliverable and write meta-level feedback explaining that grade. They also used the visualization of the feedback to discuss strengths and weaknesses of a project with a student team and provide additional guidance. One assistant also noted that she thought about how her feedback would appear in the tool and tried not to write comments that would appear inconsistent with the peer feedback for that same project. Future work should explore how seeing a visualization of the topic and opinion structure in their own feedback affects the feedback an instructor later writes.

Our interviews with the teaching assistants indicate there are additional opportunities to extend the tool to enhance course instruction. For instance, topic columns in the tool that contain many criticism icons across many projects might indicate gaps in project knowledge for the recipients or unrealistic expectations of the providers. An instructor might use the tool to identify such topic columns and prioritize the coverage of these issues in course materials or manage students’ expectations of their peers’ projects. Instructors might also use praise labels to curate examples of good project deliverables and best practices that could

be incorporated into lecture materials or linked as examples in the tool. Finally, instructors could measure progress on open-ended assignments by linking students' revisions back to planned changes identified in the intention labels assigned to the feedback.

To improve feedback review, the most commonly requested feature was the ability to attach notes to individual pieces of feedback, usually for explaining how a piece of feedback was addressed or why it wasn't addressed. Since feedback statements labeled with the same topic in the tool did not always address the same issue, students requested the ability to add keywords for each statement in the tool, allowing them to highlight subsets of icons that share keywords. Finally, the teaching assistants wanted students to be able to explain in the tool when the students disagreed with statements in their feedback or chose not to implement a specific suggestion. These suggested features indicate simple but powerful opportunities for improving the tool's usability for instructors and students alike.

6.7 LIMITATIONS AND FUTURE WORK

The results reported in this study were derived from a field study methodology. Though this method allowed us to answer questions about the use of a feedback visualization tool in an authentic classroom setting, we are unable to compare these results to students reviewing feedback with existing tools. Future controlled experiments are needed to determine how the choice of tool for feedback review affects feedback comprehension, interpretation strategy, and quality of revisions made to a project.

Students worked on the course projects in teams, and each team may have appropriated the tool in different ways (e.g., reviewing feedback as a team vs. reviewing feedback individually and discussing it later as a team). Our data set was not large enough to determine how the uses of the tool reported by students relate to different teamwork styles. Because it was a research prototype, the tool deployed in this study did not yet implement some collaborative features, such as assigning "must do" actions to specific members of a team or allowing feedback annotation by distributed team members simultaneously. How these and other possible collaborative features affects the use of a feedback visualization tool remains an open question. Because we deployed a specific feedback visualization tool in a single course, it is also possible students practiced feedback review processes that were not observed in our study. Moreover, the design of the feedback sessions in the course produced the type of feedback that is best visualized in the tool: feedback by multiple providers with different backgrounds, foci, and opinions. The use of the tool might not be as beneficial if used in courses where the feedback is generated from a few peers with similar backgrounds. The results reported in this study should be expanded by testing the use of feedback visualization

tools with different user populations, project types, and course cultures.

We see several additional directions for future work. One direction is to extend the tool to import and visualize the feedback written in different review contexts, such as comments posted in online critique communities (e.g., Reddit and Behance), in PDF documents, and in academic peer review platforms. A second direction is to extend the tool to support the *composition* of feedback, in addition to the interpretation of the resulting feedback collection. For example, future work could extend the tool to visualize the feedback as it is submitted and evaluate how showing providers the in-progress visualization affects the feedback they write (e.g., do they direct their comments to empty cells in the visualization?). Likewise, the tool could be extended to allow the content creator to visually mark the topics for which they are most interested in receiving feedback, and make these markings available to feedback providers at the onset of the composition process. A third direction is to leverage the tool to build an open data set of feedback statements annotated with topic, opinion, and intention labels. The data set could, for example, be used by instructors to find examples of feedback statements that promote action and learning for content creators and contrast these with examples of feedback that are not acted upon. Finally, the tool could be extended to support additional metadata such as the estimated time to implement a feedback statement. This metadata could be used to generate action plans based on how much time the content creator is willing to invest in revising the work.

6.8 CONTRIBUTIONS

Feedback for creative projects can be difficult to understand, especially when it involves resolving conflicting opinions, judging the credibility of suggestions, and prioritizing the issues raised across a multitude of topics. In this chapter, we reported findings for how student teams engaged with an interactive tool that visualizes the topic and opinion structure within a collection of feedback written by peers and the teaching assistants for their design projects. We found that the tool was useful for scaffolding students' processes for reviewing the feedback. Students leveraged the structure of the visualization in the tool to explore the critical statements in the feedback first (thumbs down icons), to assess the quality of their projects (ratio of praise to criticism icons across topics), and to facilitate team discussion about how to best revise their projects in response to the feedback. We also found that students were willing to create the meta-data needed to generate the visualization in the tool. Students indicated that the effort required to label the feedback was outweighed by the benefits of familiarizing themselves with the feedback and having access to the resulting visualization. Though the goal of the tool is to help students learn and apply skills for

feedback interpretation, the results of our study also revealed opportunities to use the tool to benefit instruction. These opportunities include curating examples of feedback with an appropriate balance of praise, criticism, and suggestions, identifying student projects that are succeeding (e.g., disproportionate praise) or in need of additional mentorship (e.g., disproportionate criticism), and identifying topics that need further explanation (e.g., seldom referenced in the feedback).

A paper reporting the results from the study discussed in this section is under submission to Transactions on Computer-Human Interaction 2022. This work makes three main contributions. First, we report emergent techniques and patterns of tool use that students leverage to accomplish their design goals for improving their in-progress creative works. Second, we offer insights and suggestions for utilizing interactive visualization to achieve desired outcomes for feedback review and instructional purposes. Finally, we explore how involving students in the process of visualizing feedback helps them gain deeper insights and develop strategies for reviewing feedback that extend beyond the classroom. These contributions can help future researchers, instructors, and designers of feedback exchange platforms better leverage tools for helping students make sense of large collections of feedback. The tool and findings reported in this study work towards a future in which students are taught not only how to write good feedback, but are also taught skills for effectively interpreting and acting on that feedback for their creative projects.

Chapter 7: General Discussion

The experiments presented in this dissertation have investigated how interface design choices such as displaying a quality score with comments (Chapter 4), displaying comments at different levels of detail (Chapter 5), and displaying comment metadata through an interactive visualization (Chapter 6) can affect the composition, interpretation, and usage of constructive feedback on creative and open-ended projects. In the remaining sections of this chapter, I discuss broader themes surrounding how interfaces facilitate communication between feedback providers and creators throughout the feedback exchange process. I also propose several practical recommendations and avenues of future work with respect to each of these themes.

7.1 PERSONALIZING FEEDBACK'S PRESENTATION

Because each creator has a unique process for interpreting feedback, many interventions that require a specific means of interacting with feedback inherently benefit some creators while inhibiting others. This theme was most prevalent in Chapter 5, where providers reported writing their comments in a single box was easier than separating them by rubric criterion, and in Chapter 6, where some students expressed a preference to reviewing plain text feedback over using the visualization. The findings from these studies advance the idea that personalizing how feedback is presented to creators and their feedback providers can facilitate more effective communication throughout the feedback exchange process. This dissertation did not test an exhaustive set of techniques for personalizing feedback, leaving the exploration of additional opportunities for personalization to future work. In this section, I describe three trajectories for personalizing the presentation of feedback based on a creator's goal orientation, their expertise within their domain, and their preferred modality of learning. I also propose concrete suggestions for personalization along each of these trajectories that may be implemented directly or explored further in future work.

One way feedback might be personalized is based on a creator's goal orientation. Goal orientation theory [167, 168, 169] postulates that an individual may be either primarily motivated by demonstrating (performance-oriented) or developing (mastery-oriented) competence. Based on this theory, creators with performance-related goals may benefit from constructive feedback paired with performance indicators such as scores (Chapter 4) or social comparisons [170] that allow them to assess their abilities relative to others. Creators with mastery-related goals may instead benefit more from constructive feedback paired with references or exemplars [6] that provide them with opportunities to learn from other creators.

An extension to goal-orientation theory distinguishes between approach and avoidance goals depending on whether an individual aims to improve or maintain their current levels of performance and mastery [171]. Feedback review interfaces might cater to approach-oriented creators by embedding tutorials for design concepts referenced in the feedback, helping them develop and demonstrate competence in new techniques. Avoidance-oriented creators may benefit from a checklist reminding them of best practices to follow, ensuring their prototypes always meet certain criteria regardless of the feedback they receive.

A creator's expertise might also be taken into account when personalizing the presentation of feedback. Novice creators often struggle to implement abstract and conceptual feedback [33], and when unable to interpret feedback, tend to make small localized revisions that don't significantly improve their work [125]. Novices may also have difficulty reconciling diverse perspectives present in a collection of feedback [172]. These challenges highlight opportunities for feedback review interfaces to present abstract, conflicting feedback in a way that is both concrete and actionable to novices. This might be accomplished by annotating feedback statements with information such as the area of the design prototype being referenced [7], the underlying design principle being discussed, or images showing how other creators addressed similar feedback. By contrast, expert creators typically prefer high-level feedback that lets them utilize their own experience to discern the appropriate low-level revisions [33, 173]. Feedback review interfaces tailored to experts might instead show a list of high-level revisions suggested in the feedback with note spaces for a creator to describe how (or if) they will address each problem.

One other factor interface designers might consider in deciding how to effectively personalize feedback presentation is a creator's learning style. The VARK inventory [174, 175] is commonly used to classify an individual as either a visual, auditory, read-write, or kinesthetic learner, or as some combination of the above. Creators who prefer a reading- and writing-based style of learning might benefit the most from traditional plain-text representations of feedback. This feedback might be supplemented by additional text-based interface elements for helping creators organize and interpret feedback, such as tags or topic labels. Creators inclined towards visual learning may prefer graphical representations and visualizations of the feedback they receive. These might include visual annotations mapping feedback statements to design prototype features, or bar charts displaying summaries of feedback providers' sentiments across different topics. Allowing creators to record audio notes for each piece of feedback they receive and play those notes back when they review the feedback later might be useful for auditory learning styles. Finally, creators who are kinesthetic learners may get the most out of interacting with their feedback through drag and drop mechanisms, such as with kanban boards [176].

The suggestions above highlight only a small number of factors that might be considered when personalizing feedback’s presentation. This list is by no means exhaustive, and creators may benefit from personalization with respect to other factors (e.g., whether their project is in an early vs. late stage.) Future work should explore how tailoring feedback presentation to individual creators (e.g., based on responses to preference surveys) can improve upon existing means of presenting feedback. Future work might also explore the advantages of putting presentational choices directly into the hands of creators, possibly by offering multiple feedback views and allowing creators to customize how an interface presents feedback to them.

7.2 GENERATING AND PRESENTING METADATA

The collection of topic, opinion, and provider metadata was essential to the visualization presented in Chapter 6, providing additional information that helped contextualize and facilitate feedback interpretation. The use of upvotes and scores seen in Chapters 3 and 4 lend additional support to the idea that metadata can play a central role in facilitating feedback exchange. In the following paragraphs, I briefly describe additional types and modalities of metadata that could help facilitate feedback exchange, discuss the application of this metadata to additional contexts, and explore the tradeoffs associated with using different methodologies for generating feedback metadata.

An important direction for future work would be to test how presenting types of metadata beyond those described in this dissertation influence the composition and interpretation of feedback. For example, prior work has shown cues about a provider’s effort in writing feedback can affect perceived feedback quality as much as the provider’s expertise [20]. A future study might investigate how presenting a *recipient’s* effort cues (e.g., the time they spent working on their design prototype or reviewing past feedback) influences the type of feedback a provider writes. This type of information could be used to help providers identify which creators are most receptive to feedback and adjust the time and effort they spend writing feedback accordingly. Providing creators with metadata they would otherwise deduce and record themselves may also alleviate cognitive load and help them focus on attending to the feedback. Future work might thus test how presenting information such as feedback’s genre of discourse [89], emotional arousal [177], or implementation difficulty influence a creator’s interpretation and usage of the feedback.

Deciding how a feedback exchange interface presents information can be as important as deciding what information it presents. Design choices regarding text’s size, color, and placement on the screen all have the potential to influence how text-based information is

interpreted. Numeric and categorical information such as scores or opinions offers even more flexibility in terms of presentational modalities. E.g., one might imagine representing a score using color, size, shape, position, or some combination of modalities. Additionally, the most effective modality for presenting a given type of information may vary from person to person. Future work could explore how presenting feedback and metadata in different modalities influence creators and providers throughout the feedback exchange process. As this dissertation has demonstrated, however, even if creators find individual interface elements and features useful, too much information can confuse and overwhelm creators, and at worst can hinder communication between creators and feedback providers. Future research should also explore which feedback cues and metadata are most effective when they are always visible vs. when they are revealed on-demand (e.g., through tooltips or popup menus).

The use of feedback metadata is not necessarily limited to feedback composition and interpretation, and might be valuable to other stages of feedback exchange or contexts where feedback is used. One could imagine importing the topic and opinion metadata from Chapter 6’s visualization into a document editor (e.g., Microsoft Word), allowing users to leverage features of the editor to manipulate the feedback based on its metadata. This would open possibilities such as searching and highlighting feedback by its topic, or importing all suggestions into a tabular checklist for reference when revising a design prototype. Another use case might be exporting visual markers on a design prototype corresponding to feedback [7] and importing them into a graphics editor (e.g., Photoshop). This metadata might be used to extend the graphics editor’s interface through a plugin for displaying context-sensitive feedback relative to the cursor position. Such a feature could allow graphic designers to selectively display the feedback most relevant to the area of their design that they are currently working on.

Much of the metadata discussed above is expensive and time-consuming to generate, highlighting an opportunity for future work to compare alternative methodologies for generating metadata. Such methodologies can be broadly divided into those that leave metadata generation to 1) the feedback provider, 2) the feedback recipient, 3) a third party (such as instructors or crowds), or 4) an automated system, with each type having its own unique tradeoffs. Providers who annotate their own feedback would likely produce the most accurate labels due to understanding their own state of mind when writing the feedback. However, requiring providers to both write and annotate critiques may not be an effective or desirable use of their limited time. Creators who generate the metadata for the feedback they receive may gain a deeper understanding of the feedback, but may not find the metadata as useful if the cost of generating it themselves is too high (Chapter 6). While leaving metadata generation to a third party allows providers to focus on critiquing and creators to focus

	By Provider	By Recipient	By 3rd Party Expert	By Peers / Crowds	Automated
<u>RECIPIENT PERSPECTIVE</u>					
Design Quality	C4			C3	C4
Feedback Topic	C5	C6	C6		
Feedback Opinion		C6	C6		
Feedback Urgency	C5			C5	[9]
Provider Expertise	C6		[9]		
Feedback Genre			C 5		
Example Designs	[6]				
Feedback Arousal			[9]		
Provider Effort			[20]		
Visual Annotations				[7]	
Tags				[7]	
Feedback Difficulty					
<u>PROVIDER PERSPECTIVE</u>					
Feedback Topic	C5				
Feedback Urgency	C5				
Recipient Expertise		[10]			
Recipient Effort					
Example Feedback					

Table 7.1: A map of prior work exploring the presentation of metadata generated by different parties. Green cells were explored in this dissertation (C# = Chapter #), while blue cells were explored in prior work. Boldface indicates underexplored metadata at the time of this writing. At most one study is referenced in each cell for brevity.

on using the feedback, outsourcing metadata generation could be prohibitively expensive. Finally, automatically generating metadata can be both quick and inexpensive, but current technology may not be able to accurately generate certain types of meta data (e.g., topic labels) without a human in the loop. Prior work typically considers at most one or two of these methodologies at a time, leaving the door open for several studies comparing different approaches to generating metadata (e.g., provider vs. recipient vs. AI). Table 7.1 highlights opportunities for exploring alternative techniques for collecting several types of metadata discussed above.

7.3 FACILITATING COMPOSITION IN DIFFERENT CONTEXTS

Feedback providers in the studies comprising this dissertation noted that both the choice of interface used to compose feedback (Chapter 5) and knowledge of how their feedback would be presented to the recipient (Chapter 6) influenced their writing styles. While some providers reported these interfaces helped them communicate their ideas effectively, others perceived they were primed with ideas they would not have otherwise had on their own or constrained them in some other way. These findings warrant a more in-depth examination of how different scenarios and contexts might benefit from different feedback composition interfaces. In the remainder of this section, I explore the use of several techniques for facilitating feedback composition based on a provider's time constraints, the number of providers and recipients involved in feedback exchange, and the unique challenges associated with different feedback environments.

One of the most common external factors impacting feedback composition is the issue of time constraints. Whether a provider is willing and able to spend 2 minutes, 10 minutes, or 1 hour composing feedback can considerably impact the depth and quality of feedback they produce. An instructor grading 100 projects over the weekend would unlikely be able to write multiple paragraphs of feedback on each one, whereas ticking boxes on a rubric might not be a desirable way of grading 5-10 projects over the same time frame. While choosing a composition interface based on time constraints is conceptually straightforward, providers may not know ahead of time how much effort they want or need to spend writing feedback for a particular project. One solution is to have an interface prompt for increasingly detailed feedback in phases and allow the provider to submit their feedback after any phase. For example, an interface could first ask a provider to rate a project using a slider, then offer one suggestion for improvement, followed by a bulleted list of additional suggestions, and finally an explanation for each suggestion. If providing comparable feedback to several creators is desirable (as in a classroom setting), this technique might be scaled by iterating over every creator in order of need (i.e., lowest scores first) at each phase.

On the topic of scalability, the designs of feedback composition interfaces are often optimized for one-to-one communication between a single provider and a single recipient. I.e., each instance of a composition interface is agnostic to whether a provider has previously written feedback, or to whether the target recipient has previously received feedback. This gap highlights several opportunities to extend these interfaces to support one-to-many, many-to-one, and many-to-many composition use cases. When writing feedback for multiple recipients, a provider might be shown reusable snippets and templates as they type based on similarity to previous feedback they've written. If scores are involved (as when grading a large class),

an interface might also suggest scores based on those provided the last time the provider gave similar feedback. When many providers critique a single recipient (as in design critique studios), an interface may indicate topics that have already been discussed and encourage providers to vote on existing comments while commenting only on new topics. Scenarios where many providers critique many recipients (as in peer feedback sessions) may benefit from both of the above additions. Such providers might also benefit from statistics regarding how harsh or lenient their feedback is compared to other providers', as well as cues and suggestions for moderating the tone of their feedback appropriately.

Many environments such as workplaces, review platforms, and online design communities face unique challenges that necessitate additional considerations when designing feedback composition interfaces. For instance, workplace performance evaluations can lead to high-impact decisions regarding whether an employee receives a promotion or termination. An evaluation interface might include prompts reminding an employer to only criticize factors under the employee's control [60] or requiring detailed explanations for each aspect of their performance marked as unsatisfactory. As a second example, online review platforms are often a consumer's first stop when seeking answers to questions about specific media or products. Review interfaces on these platforms might prompt a reviewer to answer any frequently-asked questions they are able to, especially those that are underanswered. Another possibility might be to ask users to rate individual aspects of a product based on lists dynamically generated from common questions. Finally, online design communities often have several users that act as both feedback providers and recipients at different times, offering unique opportunities to integrate both roles. E.g., a composition interface might show a user their own before and after drafts of a design prototype that received similar feedback, allowing them to reflect on how they solved a problem themselves when offering advice to a new recipient. Future research should explore these scenarios and others when determining the most effective means of composing feedback across a spectrum of environments and use cases.

7.4 MAKING FEEDBACK EXCHANGE MORE ACCESSIBLE

The critique forums (Chapter 3), feedback review interfaces (Chapters 4 and 5), and visualization (Chapter 6) discussed in this dissertation each required creators to use a keyboard and mouse for navigating text and graphics to interpret feedback. The generation of feedback and metadata in Chapters 5 and 6 respectively required similar mechanisms for engaging feedback content. However, creators with visual or motor impairments may have difficulty navigating large amounts of plain text feedback or interactive visualizations, making accessibility a crucial concern when helping a broad range of users compose and

interpret feedback. While modern devices have text-to-speech and similar options that make engaging with feedback *possible* for such creators, interfaces that make feedback composition and interpretation genuinely accessible are limited. Below, I discuss a few possible interface directions to improve the accessibility of feedback interpretation and composition.

Interpretation might be made more accessible by offering multimodal representations of both feedback metadata and the feedback itself [178, 179]. For example, an interactive visualization such as the one described in Chapter 6 might include a button next to each feedback statement that describes its metadata using natural language before reading the feedback out loud (e.g., “a product design expert rates the design 9/10 and offers the following suggestion...”). Such an interface could improve feedback comprehension by conveying metadata more organically than possible when using text-to-speech to read raw metadata tags. For creators with limited mobility, another multimodal option might be to leverage eye tracking [179, 180] for displaying context-sensitive popups with actions and metadata for the currently viewed feedback statement. These popups could be customized by individual users to associate actions with eye movement patterns (e.g., looking up to mark a feedback statement as “to-do” or looking down to mark it as “completed”).

One simple strategy for making composition more accessible would be to include an option for dictating feedback using a microphone and speech-to-text software. Rubric-based interfaces such as those presented in Chapter 5 might also include audio prompts that allow a user to specify the topic or design area they are critiquing. The speech-to-text software itself could be further extended, e.g., to automatically break dictated feedback into idea units and assign opinion labels based on the speaker’s cadence and tone respectively [181, 182]. To help visually impaired users generate additional types of feedback metadata, these composition interfaces could include support for gesture-based data entry mechanisms [183, 184] to complement those that require a keyboard and mouse.

By enabling multimodal interaction for both feedback composition and interpretation, the interface tweaks described above could allow a broader and more diverse range of users to enjoy the benefits of feedback exchange support tools. For this reason, future research should explore integrating techniques similar to those described above into the design of feedback support tools going forward.

7.5 GENERALIZING INSIGHTS TO DIFFERENT FEEDBACK EXCHANGE CONTEXTS

The preceding sections have discussed contributions of this dissertation from the perspective of supporting asynchronous feedback exchange on creative works. In this section, I generalize

this dissertation’s contributions to alternate formulations of feedback exchange and to contexts outside feedback exchange entirely.

While providers often write feedback independently from one another, certain environments may require providers to collaboratively produce a single feedback document, as with hiring or paper review committees. Among other challenges, these scenarios frequently require providers to delegate work and reconcile disagreements amongst themselves. Future research should explore and assess interfaces for addressing the challenges unique to collaborative feedback composition. For instance, one could imagine a collaborative feedback composition interface that leverages NLP techniques to highlight contradictory statements within a document. Such an interface might prompt the authors of these statements to clarify their points and to resolve any conflicts through discussion before continuing to write other feedback. An interface might also detect topics repeated by multiple providers throughout a feedback document and offer suggestions for merging and reorganizing the feedback as appropriate. Alternately, an interface might structure composition by having providers generate a list of topics to cover and assigning a subset of those topics to each provider, requiring them to draft feedback on their own assigned topics before viewing and editing the rest of the feedback.

Each study comprising this dissertation has worked under the assumption that feedback is provided and reviewed asynchronously in a written format. In practice, feedback is also commonly exchanged through synchronous oral critiques, in which a creator has the opportunity to engage in live dialogue with feedback providers. Insights from this dissertation might be applied to synchronous critiques by using speech-to-text software to transcribe a conversation between creators and feedback providers, then generating metadata and visualizations using any of the previously described methodologies. However, recordings of oral critiques also afford unique opportunities for capturing and presenting information such as the tone, timing, and body language associated with feedback. Creators might find this additional information valuable, e.g., for weighing feedback’s urgency based on a provider’s tone or contextualizing a provider’s comments when reviewing the feedback later. Future work should explore techniques for curating metadata unique to synchronous critiques, and should also assess creators’ perceived usefulness of interfaces that present this metadata.

Many of the insights and techniques explored throughout this dissertation may generalize to contexts outside of feedback exchange where text composition and interpretation are important. One could envision adapting Chapter 5’s interfaces for composing and viewing feedback at different levels of detail to give discussion forum users control over how others engage with their posts. For instance, a post creator might select a scaffolded composition interface for encouraging responses to touch upon certain topics, but may select a rubric with open comments when only sentiments towards those topics are desired. Topic and

opinion visualization techniques similar to those presented in Chapter 6 might be useful in the context of classroom peer evaluations. Such techniques could help instructors identify discrepancies in the evaluations a student receives from their peers, e.g., to determine whether they contributed sufficiently to a group project. These are just a few potential applications of this dissertation's contributions, and future work should explore additional applications of personalized interfaces for engaging with written feedback or other text-based content.

Chapter 8: Conclusion

Interfaces play a large role in facilitating communication between creators and feedback providers throughout the feedback exchange process. However, evaluations of these interfaces often do not consider the unique needs of both creators and their feedback providers, and may overlook how different interfaces for composing and presenting feedback suit the needs of some better than others. This thesis examined how several interfaces with different types and organizations of information impact feedback composition, interpretation, and utilization. The main contributions of this dissertation are: 1) empirically derived guidelines for incorporating summative feedback into feedback review interfaces and encouraging revision behavior for open-ended creative works, 2) a taxonomy of recommendations for both composing and presenting feedback at different levels of detail, 3) empirical data and knowledge regarding how interactive visualization techniques help facilitate feedback interpretation, and 4) insights surrounding the benefits of having creators generate their own topic and opinion labels for interpreting feedback

These contributions can guide the design and development of future feedback support tools that consider each creator's needs, goals, and constraints when engaging in feedback exchange. Additionally, these contributions provide a strong foundation which future works can build upon to further explore the benefits of personalized interfaces for composing and exploring feedback. The works presented throughout this dissertation advance towards my vision of personalized feedback composition and presentation by informing interface design decisions to improve the quality of online feedback exchange between creators and their feedback providers.

References

- [1] J. Hattie and H. Timperley, “The power of feedback,” *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.
- [2] A. A. Lipnevich and J. K. Smith, ““i really need feedback to learn:” students’ perspectives on the effectiveness of the differential feedback messages,” *Educational Assessment, Evaluation and Accountability*, vol. 21, no. 4, p. 347, 2009.
- [3] A. Tekian, C. J. Watling, T. E. Roberts, Y. Steinert, and J. Norcini, “Qualitative and quantitative feedback in the context of competency-based education,” *Medical Teacher*, vol. 39, no. 12, pp. 1245–1249, 2017, pMID: 28927332.
- [4] R. Butler, “Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance,” *British Journal of Educational Psychology*, vol. 58, no. 1, pp. 1–14, 1988.
- [5] M. Northcote, A. Williams, P. Fitzsimmons, and P. Kilgour, “Does the type of assessment feedback i give make a difference?: the impact of qualitative and quantitative assessment feedback,” in *ICERI2014 Proceedings*, ser. 7th International Conference of Education, Research and Innovation. IATED, 17-19 November, 2014 2014, pp. 4531–4540.
- [6] H. B. Kang, G. Amoako, N. Sengupta, and S. P. Dow, “Paragon: An online gallery for enhancing design feedback with visual examples,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.
- [7] A. Xu, S.-W. Huang, and B. Bailey, “Voyant: generating structured feedback on visual designs using a crowd of non-experts,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1433–1444.
- [8] M. D. Greenberg, M. W. Easterday, and E. M. Gerber, “Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers,” in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 2015, pp. 235–244.
- [9] K. Luther, J.-L. Tolentino, W. Wu, A. Pavel, B. P. Bailey, M. Agrawala, B. Hartmann, and S. P. Dow, “Structuring, aggregating, and evaluating crowdsourced design critique,” in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 473–485.
- [10] R. Cheng, Z. Zeng, M. Liu, and S. Dow, “Critique me: Exploring how creators publicly request feedback in an online critique community,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–24, 2020.

- [11] J. Cambre, S. Klemmer, and C. Kulkarni, “Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.
- [12] C. M. Hicks, V. Pandey, C. A. Fraser, and S. Klemmer, “Framing feedback: Choosing review environment features that support high quality peer assessment,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 458–469.
- [13] A. Yuan, K. Luther, M. Krause, S. I. Vennix, S. P. Dow, and B. Hartmann, “Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 1005–1017.
- [14] H. Schneider, K. Frison, J. Wagner, and A. Butz, “Crowdux: a case for using widespread and lightweight tools in the quest for ux,” in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 2016, pp. 415–426.
- [15] A. Xu, H. Rao, S. P. Dow, and B. P. Bailey, “A classroom study of using crowd feedback in the iterative design process,” in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 1637–1648.
- [16] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey, “Listen to others, listen to yourself: Combining feedback review and reflection to improve iterative design,” in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 2017, pp. 158–170.
- [17] Y. W. Wu and B. P. Bailey, “Soften the pain, increase the gain: Enhancing users’ resilience to negative valence feedback,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.
- [18] R. Butler and M. Nisan, “Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance.” *Journal of Educational Psychology*, vol. 78, no. 3, pp. 210–216, 1986.
- [19] A. A. Lipnevich and J. K. Smith, “Response to assessment feedback: the effects of grades, praise, and source of information,” *ETS Research Report Series*, vol. 2008, no. 1, pp. i–57, 2008.
- [20] Y. W. Wu and B. P. Bailey, “Novices who focused or experts who didn’t?” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4086–4097.
- [21] Y. W. Wu and B. P. Bailey, “Bitter sweet or sweet bitter? how valence order and source identity influence feedback acceptance,” in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 2017, pp. 137–147.

- [22] Y.-C. G. Yen, J. O. Kim, and B. P. Bailey, “Decipher: an interactive visualization tool for interpreting unstructured design feedback from multiple providers,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.
- [23] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, “Peer and self assessment in massive online classes,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 6, pp. 1–31, 2013.
- [24] K. Luther, A. Pavel, W. Wu, J.-I. Tolentino, M. Agrawala, B. Hartmann, and S. P. Dow, “Crowdcrit: crowdsourcing and aggregating visual design critique,” in *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 21–24.
- [25] D. Carless, “Differing perceptions in the feedback process,” *Studies in higher education*, vol. 31, no. 2, pp. 219–233, 2006.
- [26] R. Higgins, P. Hartley, and A. Skelton, “The conscientious consumer: Reconsidering the role of assessment feedback in student learning,” *Studies in higher education*, vol. 27, no. 1, pp. 53–64, 2002.
- [27] M. Price, K. Handley, J. Millar, and B. O’donovan, “Feedback: all that effort, but what is the effect?” *Assessment & Evaluation in Higher Education*, vol. 35, no. 3, pp. 277–289, 2010.
- [28] M. Jackson and L. Marks, “Improving the effectiveness of feedback by use of assessed reflections and withholding of grades,” *Assessment & Evaluation in Higher Education*, vol. 41, no. 4, pp. 532–547, 2016.
- [29] J. S. Goodman, R. E. Wood, and M. Hendrickx, “Feedback specificity, exploration, and learning.” *Journal of Applied Psychology*, vol. 89, no. 2, p. 248, 2004.
- [30] J. S. Goodman and R. E. Wood, “Feedback specificity, learning opportunities, and learning.” *Journal of Applied Psychology*, vol. 89, no. 5, p. 809, 2004.
- [31] R. G. Bing-You, J. Paterson, and M. A. Levine, “Feedback falling on deaf ears: residents’ receptivity to feedback tempered by sender credibility,” *Medical teacher*, vol. 19, no. 1, pp. 40–44, 1997.
- [32] P. A. Crain and B. P. Bailey, “Share once or share often? exploring how designers approach iteration in a large online community,” in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 2017, pp. 80–92.
- [33] E. Foong, D. Gergle, and E. M. Gerber, “Novice and expert sensemaking of crowdsourced design feedback,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–18, 2017.
- [34] N. E. Winstone, R. A. Nash, J. Rowntree, and M. Parker, “‘it’d be useful, but i wouldn’t use it’: barriers to university students’ feedback seeking and recipience,” *Studies in Higher Education*, vol. 42, no. 11, pp. 2026–2041, 2017.

- [35] K. R. Butcher and T. Sumner, “Self-directed learning and the sensemaking paradox,” *Human-Computer Interaction*, vol. 26, no. 1-2, pp. 123–159, 2011.
- [36] P. J. Hinds, “The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance.” *Journal of experimental psychology: applied*, vol. 5, no. 2, p. 205, 1999.
- [37] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld, “Organizing and the process of sense-making,” *Organization science*, vol. 16, no. 4, pp. 409–421, 2005.
- [38] M. A. Dijks, L. Brummer, and D. Kostons, “The anonymous reviewer: the relationship between perceived expertise and the perceptions of peer feedback in higher education,” *Assessment & Evaluation in Higher Education*, vol. 43, no. 8, pp. 1258–1271, 2018.
- [39] L. Ying-Leh, A. G. K. Abdullah, and A. Ismail, “Feedback environment and coaching communication in malaysia education organizations,” *Asian Journal of Social Sciences & Humanities Vol*, vol. 4, p. 1, 2015.
- [40] J. McCarthy, “Enhancing feedback in higher education: Students’ attitudes towards online and in-class formative assessment feedback models,” *Active Learning in Higher Education*, vol. 18, no. 2, pp. 127–141, 2017.
- [41] A. Cook, J. Hammer, S. Elsayed-Ali, and S. Dow, “How guiding questions facilitate feedback exchange in project-based learning,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
- [42] Y. Xu and D. Carless, “‘only true friends could be cruelly honest’: cognitive scaffolding and social-affective support in teacher feedback literacy,” *Assessment & Evaluation in Higher Education*, vol. 42, no. 7, pp. 1082–1094, 2017.
- [43] Y. Trope, B. Gervy, and N. Bolger, “The role of perceived control in overcoming defensive self-evaluation,” *Journal of Experimental Social Psychology*, vol. 39, no. 5, pp. 407–419, 2003.
- [44] R. Butler, “Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance.” *Journal of Educational Psychology*, vol. 79, no. 4, pp. 474–482, 1987.
- [45] G. Gibbs and C. Simpson, “Conditions under which assessment supports students’ learning,” *Learning and teaching in higher education*, no. 1, pp. 3–31, 2005.
- [46] Y.-C. Yen, “Scaffolding feedback interpretation process for creative work through reflection, paraphrasing, and information visualization,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2021.
- [47] A. Cook, S. Dow, and J. Hammer, “Designing interactive scaffolds to encourage reflection on peer feedback,” in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2020, pp. 1143–1153.

- [48] M. Jasim, E. Hoque, A. Sarvghad, and N. Mahyar, “Communitypulse: Facilitating community input analysis by surfacing hidden insights, reflections, and priorities,” in *Designing Interactive Systems Conference 2021*, 2021, pp. 846–863.
- [49] K. Yatani, M. Novati, A. Trusty, and K. N. Truong, “Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 1541–1550.
- [50] S. Andolina, H. Schneider, J. Chan, K. Klouche, G. Jacucci, and S. Dow, “Crowdboard: augmenting in-person idea generation with real-time crowds,” in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 2017, pp. 106–118.
- [51] M. X. Liu, J. Hsieh, N. Hahn, A. Zhou, E. Deng, S. Burley, C. Taylor, A. Kittur, and B. A. Myers, “Unakite: Scaffolding developers’ decision-making using the web,” in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019, pp. 67–80.
- [52] A. Xu and B. Bailey, “What do you think? a case study of benefit, expectation, and interaction in a large online critique community,” in *Proceedings of the acm 2012 conference on computer supported cooperative work*, 2012, pp. 295–304.
- [53] J. Marlow and L. Dabbish, “From rookie to all-star: professional development in a graphic design social networking site,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 922–933.
- [54] H. Wauck, Y.-C. Yen, W.-T. Fu, E. Gerber, S. P. Dow, and B. P. Bailey, “From in the class or in the wild? peers provide better design feedback than external crowds,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 5580–5591.
- [55] Y.-C. Yen, S. P. Dow, E. Gerber, and B. P. Bailey, “Social network, web forum, or task market? comparing different crowd genres for design feedback exchange,” in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 2016, pp. 773–784.
- [56] J. Hui, A. Glenn, R. Jue, E. Gerber, and S. Dow, “Using anonymity and communal efforts to improve quality of crowdsourced feedback,” in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- [57] A. Kluger and A. DeNisi, “The effects of feedback interventions on performance: An historical review, meta-analysis and preliminary feedback theory,” *Psychological Bulletin*, vol. 119, pp. 254–285, 1996.
- [58] D. R. Sadler, “Formative assessment and the design of instructional systems,” *Instructional Science*, vol. 18, no. 2, pp. 119–144, Jun 1989.
- [59] J. Lefroy, C. Watling, P. W. Teunissen, and P. Brand, “Guidelines: the do’s, don’ts and don’t knows of feedback for clinical education,” *Perspectives on medical education*, vol. 4, no. 6, pp. 284–299, 2015.

- [60] C. Toxtli, A. Richmond-Fuller, and S. Savage, “Reputation agent: Prompting fair reviews in gig markets,” in *Proceedings of The Web Conference 2020*, 2020, pp. 1228–1240.
- [61] Y.-H. Lai and J. S. Chang, “Tellmewhy: Learning to explain corrective feedback for second language learners,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 2019, pp. 235–240.
- [62] S. J. Adams, S. L. Adams, L. J. Pryor et al., “Customization of instant feedback for integrated assignments: A case study,” *Journal of Business Case Studies (JBCS)*, vol. 2, no. 1, pp. 11–22, 2006.
- [63] F. Chetwynd and C. Dobbyn, “Assessment, feedback and marking guides in distance education,” *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 26, no. 1, pp. 67–78, 2011.
- [64] A. Shannon, J. Hammer, H. Thurston, N. Diehl, and S. Dow, “Peerpresents: A web-based system for in-class peer feedback during student presentations,” in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 2016, pp. 447–458.
- [65] J. B. Moghadam, R. R. Choudhury, H. Yin, and A. Fox, “Autostyle: Toward coding style feedback at scale,” in *Proceedings of the Second (2015) ACM Conference on Learning Scale*, 2015, pp. 261–266.
- [66] E. L. Glassman, L. Fischer, J. Scott, and R. C. Miller, “Foobaz: Variable name feedback for student code at scale,” in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 609–617.
- [67] A. Head, E. Glassman, G. Soares, R. Suzuki, L. Figueredo, L. D’Antoni, and B. Hartmann, “Writing reusable code feedback at scale with mixed-initiative program synthesis,” in *Proceedings of the Fourth (2017) ACM Conference on Learning Scale*, 2017, pp. 89–98.
- [68] T. J. Ngoon, C. A. Fraser, A. S. Weingarten, M. Dontcheva, and S. Klemmer, “Interactive guidance techniques for improving creative feedback,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–11.
- [69] S. Mahajan, “Toward automated critique for student-created interactive narrative projects,” in *Proceedings of the... AAAI Conference on Artificial Intelligence*, 2019.
- [70] A. Bharadwaj, P. Siangliulue, A. Marcus, and K. Luther, “Critic: Augmenting creative work with dynamic checklists, automated quality assurance, and contextual reviewer feedback,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

- [71] M. Cutumisu, K. P. Blair, D. B. Chin, and D. L. Schwartz, “Assessing whether students seek constructive criticism: The design of an automated feedback system for a graphic design task,” *International Journal of Artificial Intelligence in Education*, vol. 27, no. 3, pp. 419–447, 2017.
- [72] I. Solecki, J. Porto, N. d. C. Alves, C. Gresse von Wangenheim, J. Hauck, and A. F. Borgatto, “Automated assessment of the visual design of android apps developed with app inventor,” in *Proceedings of the 51st ACM technical symposium on computer science education*, 2020, pp. 51–57.
- [73] P. W. Foltz and M. Rosenstein, “Analysis of a large-scale formative writing assessment system with automated feedback,” in *Proceedings of the Second (2015) ACM Conference on Learning Scale*, 2015, pp. 339–342.
- [74] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi, “Using learning analytics to scale the provision of personalised feedback,” *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, 2019.
- [75] A. Malik, M. Wu, V. Vasavada, J. Song, J. Mitchell, N. Goodman, and C. Piech, “Generative grading: Neural approximate parsing for automated student feedback,” *arXiv preprint arXiv:1905.09916*, 2019.
- [76] V. S. Kumar and D. Boulanger, “Automated essay scoring and the deep learning black box: How are rubric scores determined?” *International Journal of Artificial Intelligence in Education*, pp. 1–47, 2020.
- [77] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, and S. R. Klemmer, “The effect of parallel prototyping on design performance, learning, and self-efficacy,” *Stanford Technical Report*, vol. 10, 2009.
- [78] E. Gerber and M. Carroll, “The psychological experience of prototyping,” *Design studies*, vol. 33, no. 1, pp. 64–84, 2012.
- [79] J. Nielsen, “Iterative user-interface design,” *Computer*, vol. 26, no. 11, pp. 32–41, 1993.
- [80] D. A. Schon, “Designing as reflective conversation with the materials of a design situation,” *Research in engineering design*, vol. 3, no. 3, pp. 131–147, 1992.
- [81] B. P. Bailey and E. Horvitz, “What’s your idea? a case study of a grassroots innovation pipeline within a large software company,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 2065–2074.
- [82] E. Gilbert, “Widespread underprovision on reddit,” in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 803–808.
- [83] C. Lampe and P. Resnick, “Slash (dot) and burn: distributed moderation in a large online conversation space,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 543–550.

- [84] Y. J. Reimer and S. A. Douglas, “Teaching hci design with the studio approach,” *Computer science education*, vol. 13, no. 3, pp. 191–205, 2003.
- [85] J. Chan, S. P. Dow, and C. D. Schunn, “Do the best design ideas (really) come from conceptually distant sources of inspiration?” in *Engineering a Better Future*. Springer, Cham, 2018, pp. 111–139.
- [86] S. P. Dow, K. Heddlestone, and S. R. Klemmer, “The efficacy of prototyping under time constraints,” in *Proceedings of the seventh ACM conference on Creativity and cognition*, 2009, pp. 165–174.
- [87] J. Elkins, *Art critiques: A guide*. New Academia Publishing, 2012.
- [88] J. Shanteau, D. J. Weiss, R. P. Thomas, and J. C. Pounds, “Performance-based assessment of expertise: How to decide if someone is an expert or not,” *European Journal of Operational Research*, vol. 136, no. 2, pp. 253–263, 2002.
- [89] D. P. Dannels and K. N. Martin, “Critiquing critiques: A genre analysis of feedback across novice to expert design studios,” *Journal of Business and Technical Communication*, vol. 22, no. 2, pp. 135–159, 2008.
- [90] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.
- [91] A. J. Viera, J. M. Garrett et al., “Understanding interobserver agreement: the kappa statistic,” *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.
- [92] B. J. McInnis, E. L. Murnane, D. Epstein, D. Cosley, and G. Leshed, “One and done: Factors affecting one-time contributors to ad-hoc online communities,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 609–623.
- [93] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala, “Commentspace: structured support for collaborative visual analysis,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2011, pp. 3131–3140.
- [94] S. Doroudi, E. Kamar, E. Brunskill, and E. Horvitz, “Toward a learning science for complex crowdsourcing tasks,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 2623–2634.
- [95] K. Luther, K. Caine, K. Ziegler, and A. Bruckman, “Why it works (when it works) success factors in online creative collaboration,” in *Proceedings of the 16th ACM international conference on Supporting group work*, 2010, pp. 1–10.
- [96] L. Mamykina, B. Manóim, M. Mittal, G. Hripcsak, and B. Hartmann, “Design lessons from the fastest q&a site in the west,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2011, pp. 2857–2866.

- [97] J. Kim, M. Agrawala, and M. S. Bernstein, “Mosaic: designing online creative communities for sharing works-in-progress,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 246–258.
- [98] S. Dow, E. Gerber, and A. Wong, “A pilot study of using crowds in the classroom,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’13. New York, NY, USA: Association for Computing Machinery, 2013. [Online]. Available: <https://doi.org/10.1145/2470654.2470686> p. 227–236.
- [99] C. Eastman, “Automated assessment of early concept designs,” *Architectural Design*, vol. 79, no. 2, pp. 52–57, 2009.
- [100] R. Symons, “In their own words: Finding out what students think about their university learning experience,” *Synergy*, vol. 23, pp. 34–35, 2006.
- [101] D. R. Sadler, “Beyond feedback: Developing student capability in complex appraisal,” *Assessment & Evaluation in Higher Education*, vol. 35, no. 5, pp. 535–550, 2010.
- [102] F. Raczkowski, “It’s all fun and games... a history of ideas concerning gamification.” in *DiGRA Conference*, 2013.
- [103] “Amazon mechanical turk (2019),” 2019. [Online]. Available: <https://www.mturk.com/>
- [104] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, “Soylent: A word processor with a crowd inside,” in *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’10. New York, NY, USA: ACM, 2010, pp. 313–322.
- [105] J. Kim and A. Monroy-Hernández, “Storia: Summarizing social media content based on narrative theory using crowdsourcing,” *CoRR*, vol. abs/1509.03026, 2015.
- [106] M. Nebeling, A. To, A. Guo, A. A. de Freitas, J. Teevan, S. P. Dow, and J. P. Bigham, “Wearwrite: Crowd-assisted writing from smartwatches,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16. New York, NY, USA: ACM, 2016, pp. 3834–3846.
- [107] J. Teevan, S. T. Iqbal, and C. von Veh, “Supporting collaborative writing with micro-tasks,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16. New York, NY, USA: ACM, 2016, pp. 2657–2668.
- [108] “Reddit (2019),” 2019. [Online]. Available: <https://www.reddit.com/r/mturk/>
- [109] M. Vaezi and S. Rezaei, “Development of a rubric for evaluating creative writing: a multi-phase research,” *New Writing*, vol. 16, no. 3, pp. 303–317, 2019.
- [110] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

- [111] H. Brenner and U. Kliebsch, “Dependence of weighted kappa coefficients on the number of categories,” *Epidemiology*, vol. 7, no. 2, pp. 199–202, Mar 1996.
- [112] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.
- [113] S. Greenland, “Principles of multilevel modelling,” *International journal of epidemiology*, vol. 29, no. 1, pp. 158–167, 2000.
- [114] E. Pedhazur, *Multiple regression in behavioral research*. Orlando, FL: Harcourt, 1997.
- [115] M. Lewis, “Stepwise versus hierarchical regression: Pros and cons.” *Online Submission*, 2007.
- [116] R. F. Kizilcec, “How much information? effects of transparency on trust in an algorithmic interface,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 2390–2395.
- [117] C. An, “The content and role of intrinsic motivation in creative work: the importance of seeking “enjoyment”,” *Creativity Studies*, vol. 12, no. 2, pp. 280–290, 2019.
- [118] M. Benedek, R. Bruckdorfer, and E. Jauk, “Motives for creativity: Exploring the what and why of everyday creativity,” *The Journal of Creative Behavior*, vol. 54, no. 3, pp. 610–625, 2020.
- [119] P. Winne and D. Butler, “Student cognition in learning from teaching,” *International encyclopedia of education*, vol. 2, pp. 5738–5775, 1994.
- [120] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen, “Getting the right design and the design right,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’06. New York, NY, USA: ACM, 2006. [Online]. Available: <http://doi.acm.org/10.1145/1124772.1124960> pp. 1243–1252.
- [121] C. H. Edwards and L. Edwards, “Let’s end the grading game,” *The Clearing House*, vol. 72, no. 5, pp. 260–263, 1999.
- [122] D. C. Molden and C. S. Dweck, “Finding” meaning” in psychology: a lay theories approach to self-regulation, social perception, and social development.” *American Psychologist*, vol. 61, no. 3, p. 192, 2006.
- [123] B. Lane, *After the End: Teaching and Learning Creative Revision*. Heinemann, 2016.
- [124] G. A. Pignatiello, R. J. Martin, and R. L. Hickman Jr, “Decision fatigue: A conceptual analysis,” *Journal of health psychology*, p. 1359105318763510, 2018.
- [125] J. A. Butler and M. A. Britt, “Investigating instruction for improving revision of argumentative essays,” *Written Communication*, vol. 28, no. 1, pp. 70–96, 2011.
- [126] A. Sidky, J. Arthur, and S. Bohner, “A disciplined approach to adopting agile practices: the agile adoption framework,” *Innovations in systems and software engineering*, vol. 3, no. 3, pp. 203–216, 2007.

- [127] P. Black and D. Wiliam, “Assessment and classroom learning,” *Assessment in Education: principles, policy & practice*, vol. 5, no. 1, pp. 7–74, 1998.
- [128] M. Jasim, P. Khaloo, S. Wadhwa, A. X. Zhang, A. Sarvghad, and N. Mahyar, “Communityclick: Capturing and reporting community feedback from town halls to improve inclusivity,” *arXiv preprint arXiv:2009.09053*, 2020.
- [129] M. Ojha and M. Arshad Rahman, “Do online courses provide an equal educational value compared to in-person classroom teaching? evidence from us survey data using quantile regression,” *Evidence from US Survey Data Using Quantile Regression (July 14, 2020)*, 2020.
- [130] J. Yu and C. Liu, “The impact of employee participation in online innovation communities on idea quality,” *Kybernetes*, 2020.
- [131] L. Sun, Y. Tang, and W. Zuo, “Coronavirus pushes education online,” *Nature Materials*, vol. 19, no. 6, pp. 687–687, 2020.
- [132] O. R. B. Speily, A. Rezvanian, A. Ghasemzadeh, A. M. Saghiri, and S. M. Vahidipour, “Lurkers versus posters: Investigation of the participation behaviors in online learning communities,” in *Educational Networking*. Springer, 2020, pp. 269–298.
- [133] G. Gibbs, “How assessment frames student learning,” *Innovative assessment in higher education*, vol. 23, 2006.
- [134] D. Carless, D. Salter, M. Yang, and J. Lam, “Developing sustainable feedback practices,” *Studies in higher education*, vol. 36, no. 4, pp. 395–407, 2011.
- [135] K. Cho and C. MacArthur, “Student revision with peer and expert reviewing,” *Learning and instruction*, vol. 20, no. 4, pp. 328–338, 2010.
- [136] H. L. Roediger III and M. A. Pyc, “Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice,” *Journal of Applied Research in Memory and Cognition*, vol. 1, no. 4, pp. 242–248, 2012.
- [137] C. F. Staltari, A. Baft-Neff, L. J. Marra, and G. J. Rentschler, “Supervision: formative feedback for clinical documentation in a university speech-language pathology program,” *Perspectives on Administration and Supervision*, vol. 20, no. 3, pp. 117–123, 2010.
- [138] H. G. Andrade, “Teaching with rubrics: The good, the bad, and the ugly,” *College teaching*, vol. 53, no. 1, pp. 27–31, 2005.
- [139] W. J. Popham, “What’s wrong—and what’s right—with rubrics,” *Educational leadership*, vol. 55, no. 2, pp. 72–75, 1997.
- [140] A. Kohn, “The trouble with rubrics,” *English journal*, vol. 95, no. 4, pp. 12–15, 2006.
- [141] M. Wilson, *Rethinking rubrics in writing assessment*. Heinemann Portsmouth, NH, 2006.

- [142] S. Ashton and R. S. Davies, “Using scaffolded rubrics to improve peer assessment in a mooc writing course,” *Distance education*, vol. 36, no. 3, pp. 312–334, 2015.
- [143] P. Gibbons, *Scaffolding language, scaffolding learning*. Portsmouth, NH: Heinemann, 2002.
- [144] K. M. Wong and B. L. Moorhouse, “Writing for an audience: Inciting creativity among young english language bloggers through scaffolded comments,” *TESOL Journal*, vol. 9, no. 4, pp. 1–6, 2018.
- [145] M. Besser, P. Carrasco, J. Lagos, and J. Villalon, “Scaffolding feedback in writing using an online marking platform: A case study,” in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2018, pp. 277–281.
- [146] K. Halperin, C. Snyder, R. J. Shenkel, and B. K. Houston, “Effects of source status and message favorability on acceptance of personality feedback,” *Journal of Applied Psychology*, vol. 61, no. 1, p. 85, 1976.
- [147] X. Xu, J. Fan, and S. Dow, “Schema and metadata guide the collective generation of relevant and diverse work,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, no. 1, 2020, pp. 178–182.
- [148] M. Liu, Y. Li, W. Xu, and L. Liu, “Automated essay feedback generation and its impact on revision,” *IEEE Transactions on Learning Technologies*, vol. 10, no. 4, pp. 502–513, 2016.
- [149] J. G. Kim, H. K. Kong, K. Karahalios, W.-T. Fu, and H. Hong, “The power of collective endorsements: Credibility factors in medical crowdfunding campaigns,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: <https://doi.org/10.1145/2858036.2858289> p. 4538–4549.
- [150] X. Ma, L. Yu, J. L. Forlizzi, and S. P. Dow, “Exiting the design studio: Leveraging online participants for early-stage design feedback,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 676–685.
- [151] H. McGarrell and J. Verbeem, “Motivating revision of drafts through formative feedback,” *ELT journal*, vol. 61, no. 3, pp. 228–236, 2007.
- [152] Y. Kotturi and M. Kingston, “Why do designers in the” wild” wait to seek feedback until later in their design process?” in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 541–546.
- [153] A. Kasunic, C.-W. Chiang, G. Kaufman, and S. Savage, “Turker tales: Integrating tangential play into crowd work,” in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 21–34.

- [154] P. Crain and B. Bailey, “What’s the point? how scores undermine written comments on open-ended work,” in *Proceedings of the Eighth ACM Conference on Learning at Scale*, 2021, pp. 127–138.
- [155] K. Holtzblatt and H. Beyer, *Contextual design: defining customer-centered systems*. Elsevier, 1997.
- [156] I. Vardi, “The relationship between feedback and change in tertiary student writing in the disciplines,” *International Journal of Teaching and Learning in Higher Education*, vol. 20, no. 3, pp. 350–361, 2008.
- [157] C. Aitchison, “Learning from multiple voices: Feedback and authority in doctoral writing groups,” in *Writing groups for doctoral education and beyond*. Routledge, 2014, pp. 67–80.
- [158] K. Misiejuk, B. Wasson, and K. Egelandstad, “Using learning analytics to understand student perceptions of peer feedback,” *Computers in human behavior*, vol. 117, p. 106658, 2021.
- [159] N. Mahyar, M. R. James, M. M. Ng, R. A. Wu, and S. P. Dow, “Communitycrit: inviting the public to improve and evaluate urban design ideas through micro-activities,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.
- [160] M. Jasim, P. Khaloo, S. Wadhwa, A. X. Zhang, A. Sarvghad, and N. Mahyar, “Communityclick: Capturing and reporting community feedback from town halls to improve inclusivity,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–32, 2021.
- [161] D. Meulbroek, D. Ferguson, M. Ohland, and F. Berry, “Forming more effective teams using catme teammaker and the gale-shapley algorithm,” in *2019 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2019, pp. 1–5.
- [162] D. R. Thomas, “A general inductive approach for analyzing qualitative evaluation data,” *American journal of evaluation*, vol. 27, no. 2, pp. 237–246, 2006.
- [163] A. Dohrenwend, “Serving up the feedback sandwich,” *Family practice management*, vol. 9, no. 10, p. 43, 2002.
- [164] M. Kuniecki, J. Pilarczyk, and S. Wichary, “The color red attracts attention in an emotional context. an erp study,” *Frontiers in human neuroscience*, vol. 9, p. 212, 2015.
- [165] G. Morales-Martinez, P. Latreille, and P. Denny, “Nationality and gender biases in multicultural online learning environments: The effects of anonymity,” in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.
- [166] M. Jern, “Information drill-down using web tools,” in *Visualization in Scientific Computing’97*. Springer, 1997, pp. 9–20.

- [167] S. B. Button, J. E. Mathieu, and D. M. Zajac, "Goal orientation in organizational research: A conceptual and empirical foundation," *Organizational behavior and human decision processes*, vol. 67, no. 1, pp. 26–48, 1996.
- [168] D. VandeWalle, W. L. Cron, and J. W. Slocum Jr, "The role of goal orientation following performance feedback," *Journal of applied psychology*, vol. 86, no. 4, p. 629, 2001.
- [169] C. J. Waples, "Receptivity to feedback: an investigation of the influence of feedback sign, feedback specificity, and goal orientation," Ph.D. dissertation, Kansas State University, 2015.
- [170] N. W. Van Yperen, V. Brenninkmeijer, and A. P. Buunk, "People's responses to upward and downward social comparisons: The role of the individual's effort-performance expectancy," *British Journal of Social Psychology*, vol. 45, no. 3, pp. 519–533, 2006.
- [171] M. Smith, J. Duda, J. Allen, and H. Hall, "Contemporary measures of approach and avoidance goal orientations: Similarities and differences," *British Journal of Educational Psychology*, vol. 72, no. 2, pp. 155–190, 2002.
- [172] E. Foong, S. P. Dow, B. P. Bailey, and E. M. Gerber, "Online feedback exchange: A framework for understanding the socio-psychological factors," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 4454–4467.
- [173] R. E. Mayer, "From novice to expert," in *Handbook of human-computer interaction*. Elsevier, 1997, pp. 781–795.
- [174] N. D. Fleming, *Teaching and learning styles: VARK strategies*. Neil Fleming, 2001.
- [175] W. L. Leite, M. Svinicki, and Y. Shi, "Attempted validation of the scores of the fark: Learning styles inventory with multitrait-multimethod confirmatory factor analysis models," *Educational and psychological measurement*, vol. 70, no. 2, pp. 323–339, 2010.
- [176] M. Burrows, *Kanban from the Inside*. Blue Hole Press Sequim, WA, USA, 2014.
- [177] M. R. Islam and M. F. Zibran, "Deva: sensing emotions in the valence arousal space in software engineering text," in *Proceedings of the 33rd annual ACM symposium on applied computing*, 2018, pp. 1536–1543.
- [178] M. Pous, C. Serra-Vallmitjana, R. Giménez, M. Torrent-Moreno, and D. Boix, "Enhancing accessibility: Mobile to atm case study," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 2012, pp. 404–408.
- [179] J. V. Singh and G. Prasad, "Enhancing an eye-tracker based human-computer interface with multi-modal accessibility applied for text entry," *International Journal of Computer Applications*, vol. 130, no. 16, pp. 16–22, 2015.

- [180] A. Chetcuti and C. Porter, “Butterfleye: supporting the development of accessible web applications for users with severe motor-impairment,” in *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*, 2016, pp. 1–3.
- [181] Y. Xu, “Prosodypro—a tool for large-scale systematic prosody analysis.” Laboratoire Parole et Langage, France, 2013.
- [182] B. Bigi and D. Hirst, “Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody,” in *Speech Prosody*, 2012, pp. 19–22.
- [183] D. J. Ward, A. F. Blackwell, and D. J. MacKay, “Dasher: A gesture-driven data entry interface for mobile computing,” *Human–Computer Interaction*, vol. 17, no. 2-3, pp. 199–228, 2002.
- [184] M. Bennett, K. McCarthy, S. O’modhrain, and B. Smyth, “Simpleflow: enhancing gestural interaction with gesture prediction, abbreviation and autocompletion,” in *IFIP Conference on Human-Computer Interaction*. Springer, 2011, pp. 591–608.