Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples

Hyeonsu B. Kang[1] **Gabriel Amoako**[2] **Neil Sengupta**[1] **Steven P. Dow**[3]

[1]Computer Science & Engineering
[3]Cognitive Science
UC San Diego
La Jolla, CA, USA
{hyk149, nesengup, spdow}@ucsd.edu

[2]Cognitive Science & Philosophy
University of Toronto
Toronto, ON, Canada
gabriel.amoako@mail.utoronto.ca

ABSTRACT

Examples provide a source of inspiration for creating designs, but can they help improve the feedback process? Supplementing design feedback with examples could help recipients see issues clearly, identify concrete steps for improvement, and integrate novel ideas. Two online studies investigated how to support novices providing feedback on visual poster designs in an online context. Study One found that feedback providers select poster examples that complement their feedback and align with a provided rubric. Study Two shows that feedback providers give more specific, actionable, and novel input when using an example-centric approach, as opposed to text alone. To support this, we designed *Paragon*, an interface to efficiently browse examples using metadata. Finally, we discuss implications for collecting examples from the Web and structuring the design feedback process.

Author Keywords

Design; Feedback; Critique; Crowdsourcing; Examples; Interaction techniques

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Designers often collect examples in early stages of the design process to form ideas and sketch a path forward [32, 18]. Analogical transfer suggests that comparison of structurally similar problems promotes transfer from concrete comparisons to abstract analogies [11, 12, 13]. According to this theory, examples, if used effectively, may support cognition in creative work such as drawing [25, 52], ideation [43], and writing [46]. Examples can lower the barrier to entry for non-designers [28, 17]. In programming, the "opportunistic" approach emphasizes leveraging existing Web resources and snippets of program code for the benefit of "just-in-time learning of new skills and approaches," "clarifying and extending their existing knowledge," and "reminding themselves of details" [3].

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

 ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3174180

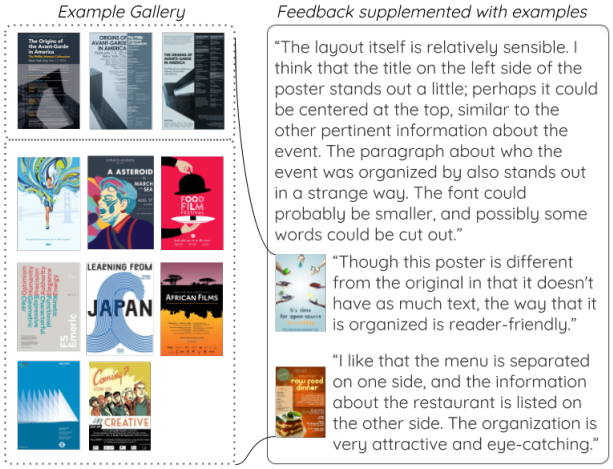

Figure 1. An overview of the example-supplemented feedback-generation process. Peer-created and Web-collected design examples with browsable crowdsourced metadata are used by feedback providers to supplement their written feedback.

While examples can be useful, novice designers often struggle to find good examples and to understand how features of examples apply to their design. Experts can help novices find examples and explain why they are relevant. In design critique – a central element of design education – designers often share work with their peers and instructors in face-to-face settings and engage in reflective dialogue [37, 15]. However, as demand for design education scales up, access to expert feedback continues to be in short supply. A number of researchers have offered tools to help peers or online crowds provide good feedback [26, 53, 29, 24]. Feedback tools have mostly focused on the speed, bandwidth, or the perceived quality of written feedback, but this paper focuses on how to support novice feedback providers by offering an online gallery of examples.

This paper presents the result of two studies investigating the value of examples as feedback components. Study One examines the efficacy and role of examples in providing design feedback through a randomized experiment with online feedback providers ($N = 30$) and in-person observation with three

participants with prior design experience. Both groups of participants incorporated examples into feedback for the provided design and found them useful. Though online participants preferred content-similar examples to diverse, Web-collected examples when providing feedback, we observed no difference in the number of references made to both visually and content-wise diverse examples. Finally, we report how feedback providers consumed examples in providing feedback, its effect on rubric use, and how examples supplemented feedback. In sum, Study One shows the efficacy of using both peer-created and diverse Web-collected examples in the design feedback process and its nuances with respect to expert-created rubric.

Study Two shows that using examples to supplement feedback, as opposed to only writing text, online feedback providers give significantly more specific, actionable, and novel input ($N = 124$) as rated by independent judges. The feedback was considered to be more descriptive, justified, and informative when supplemented by examples. Also, experts thought example-supplemented feedback provided different perspectives and insights that text-only feedback missed.

Participants in Study Two chose examples from a corpus of 287 poster designs collected from various sources to supplement their feedback. Even with our relatively small corpus, sifting through every design would be tedious for feedback providers. To understand how feedback providers accessed examples in the process of design feedback and help them efficiently find useful ones, we crowdsourced relevant metadata and implemented a filter interface. The reception of the interface was mixed; some thought it was useful for narrowing down the number of available options, but others thought it also added an additional complexity to the task. While many participants asked for a filter mechanism, designing a useful one requires striking a balance between utility and usability.

The contributions of our research are:

1. Two randomized experiments on the Web that show the efficacy, role, and benefit of using examples to supplement written feedback in the domain of visual poster design. Online novice feedback providers preferred using examples when providing feedback, and their feedback was judged by experts to be more specific, actionable, and novel compared to text-only feedback.

2. A novel interface called *Paragon* for writing online feedback supplemented by examples, along with design implications for future feedback systems.

RELATED WORK

Sources and Qualities of Effective Feedback

In design education, feedback is typically shared in co-located studio settings. Participants in design crits often share their creative work with peers, and engage in reflective communication and feedback exchange that evaluates the design and its approach, interprets the concept or artifact, and brainstorms future possibilities [6]. During a design crit, students have the benefit of sharing context and "situational awareness" [41], which provides a personalized and multi-faceted learning experience. Learners are not only exposed to better models of effective approaches [34], but also weaker ones that can serve as cautionary examples. However, design critique is hard to provide for many people, and as the demand for design education climbs, this challenge becomes even more acute.

As a way of collecting more affordable and timely feedback than in-person design crits, a number of sites offer online feedback [54, 30, 5]. However, these communities also suffer from the low quantity and reciprocity of feedback among their users [48]. Crowdsourcing feedback from novice crowds provides an asynchronous alternative to design critique and yields more feedback than existing online communities [29, 49, 16, 53]. Luther et al. have shown that conceptual, specific, and actionable feedback from online crowds can lead to higher quality designs judged by expert designers [29]. Yuan et al. found that justifications for feedback provided by novices tend to be shallow and irrelevant to their respective issues or suggestions [53]. To fill this gap, they showed that expert-created rubrics can be a scaffold for the feedback process for novices.

In addition, many researchers have studied the features that make written feedback more effective. In the domain of writing, feedback was more likely to be implemented if the problem being described was understood [31]. For visual designs, Krause et al. demonstrated the correlation between certain linguistic features of the feedback and its perceived helpfulness. Based on these features, they constructed a style guide with automatically retrieved (written feedback) examples and showed that it provides support for feedback providers [24]. Building upon this prior work, our research focuses on expanding the form of feedback by supplementing it with visual design examples. We first review relevant theories in design and cognition in the following subsection.

Theories About the Benefits of Examples

Boden describes creativity as the "sudden interlocking of two previously unrelated skills, or matrices of thought" [2]. Bringing in diverse perspectives and inspiration through examples may help designers see new possibilities that were previously unavailable. Recognizing and weighing alternatives provides value through the design process [44]. Sharing multiple designs increases not only the quality of final artifacts but also other favorable qualities, such as exploration and group rapport [7, 8].

Another possible explanation of examples' benefit may be a matter of representation. Dual-coding theory [35] would predict that examples would complement expert-created rubrics and facilitate a deeper understanding of design principles among feedback providers. Pairing multiple representations also facilitates knowledge transfer [13]; a good analogy both reveals common structure between two situations and suggests further inferences [11].

Using Examples as Feedback

Examples are pervasive in web design [28, 27, 38], prototyping [17], drawing [25], animation and video games [19], and ideation [43]. Designers use examples to support their

creative design process [18]. "Critique by redesign" in the context of data visualization can be useful [9], as Edward Tufte demonstrated in critiquing the original O-ring erosion index chart of the space shuttle Challenger [45], but it is not always possible nor appropriate (in educational settings) to take the complete-redesign approach. Adopting existing examples can be a reasonable proxy of critique-by-redesign without having to invest in full redesign.

The visual summary generation system by Robb et al. explored an approach to use examples in providing feedback [39]. In this system, they crowdsourced a visual summary consisting of a small set of representative images as feedback. Our approach differs in that we investigate the value of both written feedback and supplementary design examples instead of an *example-only* approach such as Robb et al.'s, which excludes written justification for the chosen examples.

A number of challenges arise when considering how to incorporate examples with feedback, too: (1) Examples need to be found and sourced, which can be difficult in many domains. (2) Assuming an adequate resource of examples, one needs to provide an interface that helps feedback providers find relevant examples. (3) Specifically, examples would need metadata to support search and browse. Collecting image labels through a crowdsourcing game [47], or subjective impressions with paid online workers [49] is possible. Yet, this kind of crowdwork is insufficient for our purpose as useful examples for design feedback requires deeper levels of understanding than surface-level features: crowdsourcing labels can tell us whether a poster has an elephant (surface feature), but not if the elephant provides an effective point of entry (deep feature).

For searching and filtering, Yee et al. showed that using hierarchical faceted metadata and query preview improves the usability of image search [50]. Others proposed that leveraging collaborative intelligence can be beneficial to filtering the most relevant information [14]. In contrast to these structured search or filtration processes, André et al. proposed better ways to support serendipitous discoveries [1]. Drawing from this prior work and child development theories in relational similarity of how children recognize higher-order relational matches only when they are supported by lower-order commonalities [23], we hypothesize that making a distinction between low- and high-level metadata features in a filter interface will benefit feedback providers. We report its use and design implications in Study Two.

RESEARCH QUESTIONS

This paper explores how examples affect the process of design feedback when designers and online novice feedback providers are separated. We investigate the following questions:

RQ1. Can novice feedback providers use both peer-created and Web-collected diverse examples to provide feedback?

RQ2. Is feedback supplemented with examples considered more specific, actionable, and novel by expert judges?

RQ3. Is supplementing feedback with examples attainable?

Figure 2. One of the 3 target designs (left), and an example of Content-similar (middle) and Curated (right) example designs.

RQ4. Can novice feedback providers find useful examples more efficiently with *Paragon*, a filter interface with crowdsourced design metadata?

STUDY 1: HOW FEEDBACK PROVIDERS USE EXAMPLES

Study Design

To examine how examples are used as part of a feedback process, we designed a within-subject study with three conditions: Control, Content-similar, and Curated. In Control, participants were given only a target design to critique. In Content-similar and Curated (fig. 2), participants were given a gallery of 24 examples, in addition to a target design. Presentation order of experiment conditions was counterbalanced for gender and expertise and participants were randomly assigned to an order. The two example conditions, Content-similar and Curated, were inspired from four quadrants of example types divided by content and visual similarity axes (fig. 3). Regions in the quadrants can be used to characterize how examples are accessed and used in the design process. For example, both visually and content-wise diverse examples can be curated via Web, either through serendipitous discoveries or curated searches. The Web may also be used to collect content-similar (but visually diverse) examples by leveraging search engine. Designs that share the content but exhibit different design approaches can be useful as a source of inspiration for design re-targeting. Furthermore, traditional design critiques may have created a similar evnrionment to some extent, as peer-created examples share subject matter and content but potentially adopt different approaches.

In Study One, 24 content-similar design examples were collected from a North American university-level design class, where students each submitted a poster design for a lecture series as an assignment. In addition, 24 Curated examples were collected from the Web by searching "poster design" on Google. There was no constraint in this search other than the resolution and the form factor (i.e. poster designs with vertical orientation). Though the diversity of designs was not explicitly computed, each was visibly different from the rest and unique in its subject matter. The three target designs were randomly chosen from the lower-half of the class submissions; poorer design would likely benefit most from feedback.

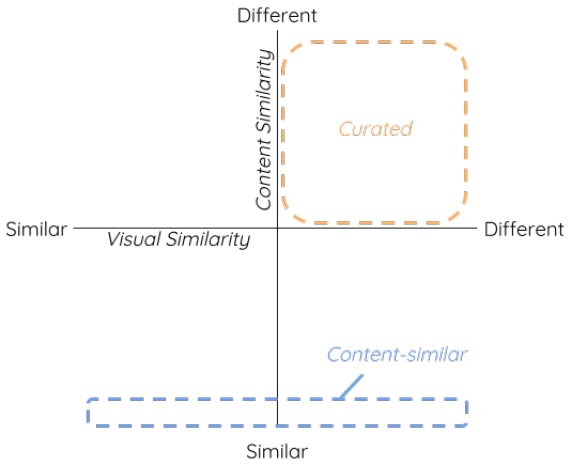

Figure 3. Quadrants of design examples. The content similarity axis represents the degree of similarity in terms of the content of designs to that of the target design. The visual similarity axis represents the degree of similarity in terms of designs' visuals.

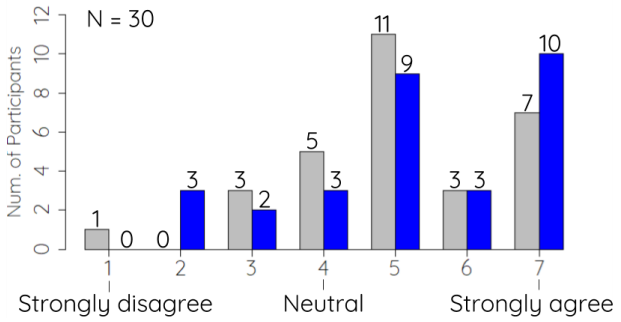

Figure 4. Responses to *"I'd like to choose an example when providing feedback"* (grey) and *"I'd like to choose similar examples rather than diverse ones when giving feedback"* (blue) show preference for choosing content-similar examples when providing feedback

Participants

32 online participants (14 female) were recruited from Amazon Mechanical Turk (MTurk). Data from two participants were removed due to failure to follow instructions. Participants' median age was 32.5 and participation was limited to U.S. residents with college education to ensure good command of written English. Two participants (6.7% of all participants) were classified as professional designers (i.e. participants who have a design degree and professional work experience or without a design degree but more than two years of work experience as a designer). Participants were compensated $4/task (∼$9.50/hr). In addition to the MTurk participants, a professional visual designer was recruited online to grade feedback quality. The designer spent three hours in total to complete the grading and was compensated $25/hr. Three university students with design work experience were recruited for in-person observation using the same task as online participants.

Apparatus

Participants were given access to an online interface (built as a Node.js Web application) to provide feedback. The interface was equipped with an image of the target design and its explanation, a gallery of browsable examples, input boxes to leave comments, and expert-created rubrics (adapted from [53]) as follows: *1. Need to consider audience 2. Provide better visual focus 3. Too much information 4. Create a more sensible layout 5. Use complementary visuals and text 6. Needs a clear visual hierarchy 7. Thoughtfully choose the typeface and colors 8. Other*. Curated and content-similar conditions differed only in terms of the examples populated in the gallery. The presentation order among examples was randomized. In the Control condition, no gallery was provided.

Procedure

Participants were given a tour of the system followed by two 1-minute training tasks (one with examples and one without examples) using the same interface. The designs used in the training tasks were collected from the Web and did not appear in subsequent tasks. After the training tasks, participants did three 5-minute tasks that corresponded to the three conditions (Control, Content-similar, and Curated). At the end of each task was a short questionnaire that asked participants the understandability of the rubric, the effectiveness in expressing their intent in the feedback, and their efficiency and ease of providing feedback. Additionally, three short, open-ended questions were administered asking the participants about their preferred qualities for examples, the way they searched for examples, and features they wished to have while searching for examples. Finally, participants were asked whether they liked to choose examples and if they preferred content-similar ones to diverse ones. The entire task took around 25 minutes.

A professional designer then graded the collected feedback. The designer was blind to the condition, but briefed that some of the feedback made references to examples and others did not. The presentation order was randomized. The designer viewed the designs and rubrics referred to in the feedback. In addition, we had in-person observations with three university students with design work experience. Each student performed the same task as online participants. Speak-aloud was used during the observations with end-of-observation debriefing.

STUDY 1: RESULTS

How did feedback providers consume examples?

Participants expressed their perception of how examples facilitated feedback generation. Some mentioned how examples aided generation of new ideas for critique by showing good 'elements':

> *"Having examples helped me think of further critique and cite designs that contained good elements that I was looking for or trying to describe."* – P12

On the other hand, the visual nature of examples itself was considered as a useful feature to "show and tell" the feedback:

> *"It was sometimes hard to put into words what I was looking for ... examples helped out a lot because they allowed me to show what I was thinking about ..."* – P26

On a 7-point Likert-scale, participants tended to prefer to choose examples when providing feedback (M = 5.1, SD = 1.48. Fig. 4, grey) compared to providing feedback without

such examples. Although only 21 participants made at least one explicit reference to examples in their feedback, examples may have helped those who did not make references in different ways. Two comments reflect this hypothesis:

> *"They (examples) displayed good qualities specific to the point on the rubric I was referencing."* – P21

While another participant commented on how examples shape the framework used to judge the quality of design:

> *"I didn't specifically refer to any examples in the feedback, but I did look at them and used them as the basis for my judgements of this flyer."* – P19

Among the participants who made at least one reference to examples, the average number of references was 1.8 in both example conditions (SD = 1.34 in Content-similar and 1.51 in Curated). Participants expressed preference for content-similar examples (M = 5.2, SD = 1.65. Fig. 4, blue) over diverse, curated ones (e.g. P18: "*When the content didn't match it was harder to explain why the example worked better*"), but the number of references to feedback made by each participant between the two example conditions was highly correlated ($r = .70$), suggesting that novice feedback providers can make a similar number of references even when the gallery contained designs both content-wise and visually diverse examples.

The effect of examples on how rubrics are used

The first and third author independently encoded rubric items in 90 (30 participants $\times$ 3 conditions) pieces of feedback. Once the initial encoding was done, the inter-rater reliability score (Cohen's κ) was first computed separately for 8 rubric items, then aggregated by weighting the average frequency of each rubric item across the encoders. The result (0.57) showed a relatively high level of agreement. On average, each of the 8 rubric items was referenced 16.5 times in Control (SD = 5.66), 15.4 times in Content-similar (SD = 6.35), and 16.1 times in Curated (SD = 5.84). On the participant's side, each of the 30 participants applied an average of 4.4, 4.1, and 4.3 (Control, Content-similar, and Curated, respectively) rubric items in their feedback.

A 3 (example type) by 6 (presentation order) repeated measures (RM) ANOVA showed no significant main or interaction effect of the presentation order on dependent variables. Therefore, we omit reporting the effect of presentation order from further discussion. We hypothesized that the distribution of applied rubrics might change as participants with examples are able to recognize previously unseen design issues. However, there was no significant effect with experiment conditions (i.e. types of example) on the number of applied rubrics nor pairwise distributional differences (chi-squared and two-sample Kolmogorov-Smirnov tests).

A two-way RM ANOVA did not show any significant main nor interaction effect of the number of rubric items and references used in the feedback on its quality score, but the trend suggests that the number of rubric items used affected its quality score ($p = 0.10$). In terms of the perceived quality of feedback, one-way RM ANCOVA did not show any significant effect of the example type ($F = .29, p = .75$). However, the trend suggests

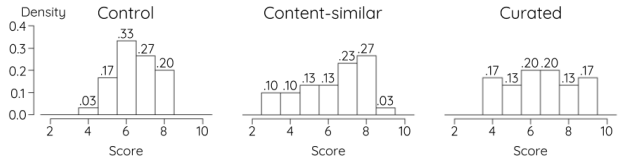

Figure 5. The distribution of perceived quality scores (in density) by different example types. Content-similar shows a more positively-skewed distribution while Curated shows a more uniform one. Difference among the distributions was suggested.

that the example type may have affected the distribution of quality scores ($\chi^2(12) = 20.0, p = .063$; fig. 5); the quality score distribution is more positively-skewed in Content-similar than Control, while Curated has the highest density of highest quality scores (the percentage of feedback that scored 9 was 17% in Curated, 3% in Content-similar and 0% in Control). The mean quality score was 6.5 (SD = 1.70) in Curated, 6.2 (SD = 1.76) in Content-similar, and 6.4 (SD = 1.10) in Control.

How to supplement feedback with examples

We categorized common criteria for choosing examples (table 1) from participants' responses to open-ended questions. We compare the frequency of these criteria between two example conditions below. 15 (50%) participants in Content-similar and 20 (67%) in Curated described their choice of examples as improvement-oriented in either visual or rubric-related qualities (e.g. P3: "*The qualities of the examples wear[sic; were] greater design layout, clarity, complementary visuals, and better visual hierarchy. Overall better use of white space.*"), meaning that they sought visual or organizational appeals from design examples that the target design lacked.

On the other hand, the number of participants who mentioned *content-similarity* and better design in *specific parts* of the design flipped between the two conditions. Four participants in Curated mentioned content-similarity as their primary factor for choosing an example, reflecting the difficulty novices face when comparing content-wise largely dissimilar designs (e.g. P19: "*(the examples) didn't provide that much information, so it's kind of hard to compare them.*"). However, only one participant in the Content-similar condition mentioned it as a criterion. This is expected, as the Content-similar gallery contained peer-created designs with shared content while the Curated gallery contained both peer-created and Web-curated diverse examples. In addition, four participants in Content-similar mentioned that they focused on specific portions (e.g. P5: *"The dates and session numbers were clearly distinguishable in the example compared to the poster I was evaluating*") that example designs did better, while only one participant in Curated mentioned a similar criterion. Mapping design elements may be harder for novices when both the content and design differ.

Observing three university students with design work experience revealed interesting comparisons with these results (fig. 6). Like novice participants, the students seeked examples that have certain qualities, whether overall visual attractiveness or things related to specific concepts that provided rubrics convey. In addition, they also mentioned compartmentalization of qualities in designs as an approach to finding useful

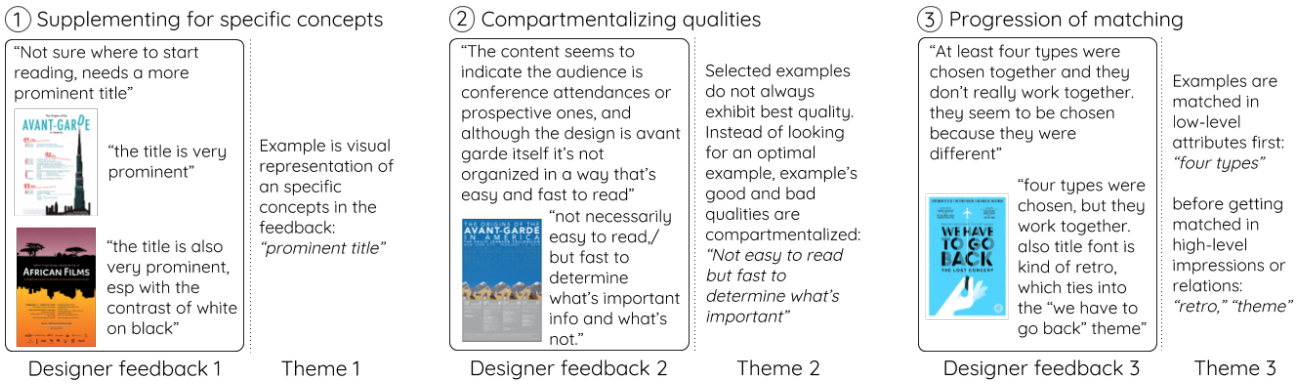

Figure 6. Three common themes appeared from in-person observation of 3 participants with prior design experience: individual concept (left), compartmentalization (middle), and progression (right).

Criteria	Content-similar	Curated
Visual Attractiveness	5 (17%)	9 (30%)
Rubric Qualities	10 (33%)	11 (37%)
Content-similarity	1 (3%)	4 (13%)
Specific (Design) Parts	4 (13%)	0 (0%)
N/A	10 (33%)	6 (20%)

Table 1. Novice feedback providers' criteria for choosing examples

examples. However, the difference may be that more experienced students do not search for the "best quality" (if any) in the gallery but rather designs with relevance and points for improvement. In order to retain relevance while comparing diverse examples, progressive matching was used: design attributes and constraints were matched, then came higher-level concepts. For example, in order to construct her feedback using diverse examples, a student first looked for designs that *used four kinds of fonts* (a low-level design attribute). With the resulting subset of designs, this student then proceeded to evaluate each design's *coherence* (a high-level concept).

STUDY 1: DISCUSSION

Study One shows that online participants used both peer-created, content-similar design examples and diverse ones on the Web to provide feedback. Novice feedback providers expressed how examples helped them see issues and express their ideas for improvement clearly. They not only tended to prefer to supplement feedback with examples, but also mentioned that examples gave them further ideas for feedback while helping them effectively communicate ideas.

Examples may also have helped novice feedback providers use the expert-created rubrics more effectively. On the one hand, feedback providers mentioned examples' role as a positive or negative reference. On the other hand, they mentioned how examples helped them attain a broader framework for understanding design principles. However, we observed no significant effect of the extent of usage and type of examples on how rubrics were used.

Online novice feedback providers chose examples that exhibited qualities they thought the target design lacked. They mentioned how qualities in design examples were compartmentalized. However, this may have been easier with peer-created examples with shared content. Compared to feedback providers with design work experience, novice participants expressed difficulty using examples that differ both in content and design as feedback components.

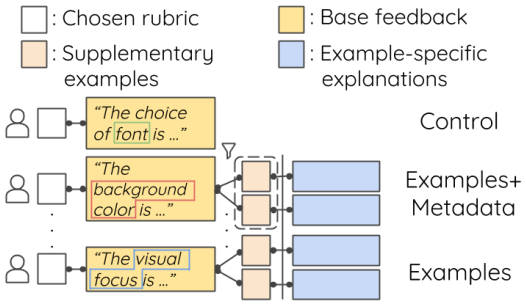

Figure 7. Participants in Study Two chooses an individual rubric item and adapts it to provide each piece of feedback. Once they finish their general feedback, they choose supplementary examples (in Examples and Examples+Metadata conditions) in stage one (left of the solid line), and then provide additional explanation specific to each supplementary example in stage two (right of the solid line). In Examples+Metadata, participants used filters to find useful examples.

Although Study One reports qualitative insights into how novice feedback providers used examples as feedback components, it did not show a significant effect of the use of examples on feedback quality. One potential reason is the effect of supplementary examples may have been confounded with that of rubrics. To disentangle these effects, we designed Study Two with a new interface and divided the feedback process into two stages: feedback providers first choose a specific rubric item to write their general feedback, then they provide an additional explanation specific to each supplementary example (fig. 7).

STUDY 2: EXAMPLES' EFFECT ON FEEDBACK QUALITY

Study design

To quantify the effect of examples on quality scores, we designed a between-subject study with three experiment conditions (Control, Examples, and Examples+Metadata) and two target designs. Each participant was randomly assigned to one of the six combinations. In Control, participants provided feedback without examples. In Examples and Examples+Metadata, participants provided feedback supplemented with examples.

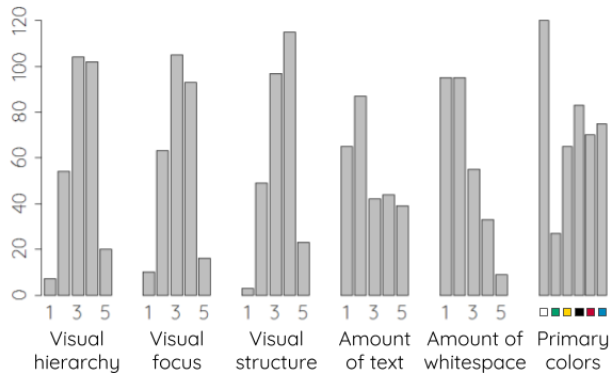

Figure 8. Distribution of 287 examples across six metadata dimensions. From left to right, *the effectiveness of visual hierarchy, focus, structure, the amount of text, white space*, and *primary colors*. Each example was rated by at least 3 people and aggregated as average scores (in 5 dimensions) or union (in Primary Color).

In Examples+Metadata, participants used filters enabled by crowdsourced metadata to help their search of examples. In addition to the target design used in Study One, an extra design was randomly selected from a set of 14 advertisement posters for a music festival [29] and used as a second target design.

Participants

Feedback providers and graders

124 online participants (69 female) were recruited from MTurk to provide feedback. Among them, 19 (15.3%) were classified as professional designers. Participants' expertise was counterbalanced and randomized. Participants were U.S. residents and had a median age of 30.5. Compensation was $1/task ($\sim$ $7.9/hr) and an extra $.3 was advertised as reward for high-quality feedback. In addition to online participants, four design experts were recruited at the university to grade the collected feedback.

Generating metadata

151 online participants (72 female, 2 other) were recruited from MTurk to generate six types of metadata about 287 design examples. The participants were U.S. residents and had a median age of 33. Compensation was $.5/task ($\sim$ $7.1/hr). Designs were randomly assigned to 9 groups of at least 31 examples for consistency. Each design was rated by at least three different participants in each dimension and assigned with the average score for scale values and the union of choices for primary colors. The dimensions for metadata-generation were four low-level attributes (the amount of text, white-space, primary colors, and content-alignment) and three interpretive qualities relevant to design principles included in the rubric (the effectiveness of visual hierarchy, focus, and structure). The content-alignment filter was pre-made; 28 examples for target one and 13 for target two. The distribution of examples is in fig. 8. To emulate an easily accessible example corpus, no specific distribution was intended in the collection phase. Participants could apply multiple filters at once.

Implementation of Paragon

Paragon is a Node.js Web application for providing feedback with online participants (fig. 9). Reduced versions were used in the Control and Examples conditions.

Procedure

After filling out consent and demographics forms, participants of the feedback-generation task were given a short tour of Paragon. The task for the example conditions had two stages. In the first stage, participants provided rubric-specific feedback and selected relevant examples. Next, they described why they chose examples (fig. 7). At the end of each task was a short questionnaire that asked participants the fairness of compensation, the easiness of finding useful examples (in Examples and Examples+Metadata), features that participants wished to have for finding useful examples (in Examples), and the usefulness of filters (in Examples+Metadata).

In the gallery, 29 out of 287 designs were peer-created and had the same content as the first target design. 13 and 18 posters were collected from two sets of designs submitted to design contests for a music festival and a lecture series, respectively [29]. The rest of designs were collected from the Web. Similar to how examples were sourced in Study One, resolution and form factor (i.e. poster designs with vertical orientation) were used as constraints for collecting the first 227 non-overlapping examples using Google image search. For the metadata-generation task, a short guideline for each dimension was provided in the beginning.

Measures

Study Two differs from Study One in the operationalization of perceived quality of feedback. Instead of an aggregated score of goodness, we used three prominent dimensions of quality (Specific, Actionable, and Novel, adapted from Sadler [40]). The Conceptual (i.e. "possess a concept of the standard – or goal, or reference level – being aimed for" [40]) dimension was replaced with Novel to capture the value of examples as a source for inspiration. Since the same expert-created rubrics were used in all conditions, the effect of examples in the Conceptual dimension is expected to vary little.

Four experts graded feedback in comparison. We chose this format instead of showing one piece of feedback at a time to discover the comparative benefit of supplementing feedback with examples. Three kinds of pairs: (1) No-example vs no-example (i.e. feedback without references to examples), (2) example vs example (i.e. feedback with references to examples), (3) and no-example vs example were presented in a random order, and the left- or right-placement in each pair was also randomized. Experts were blind to the condition in which feedback was generated. Feedback was graded on a scale of 4, from 1 = (feedback on the left is) greatly better, 2 = slightly better, 3 = (feedback on the right is) slightly better, to 4 = greatly better. Since how feedback is paired can affect the quality scores, we generated all possible pairs and randomly selected a subset for grading. All generated pairs satisfied the following conditions: (1) Both pieces of feedback are for the same target design and used the same rubric item. (2) Each piece of feedback appears at most once among all pairs. (3)

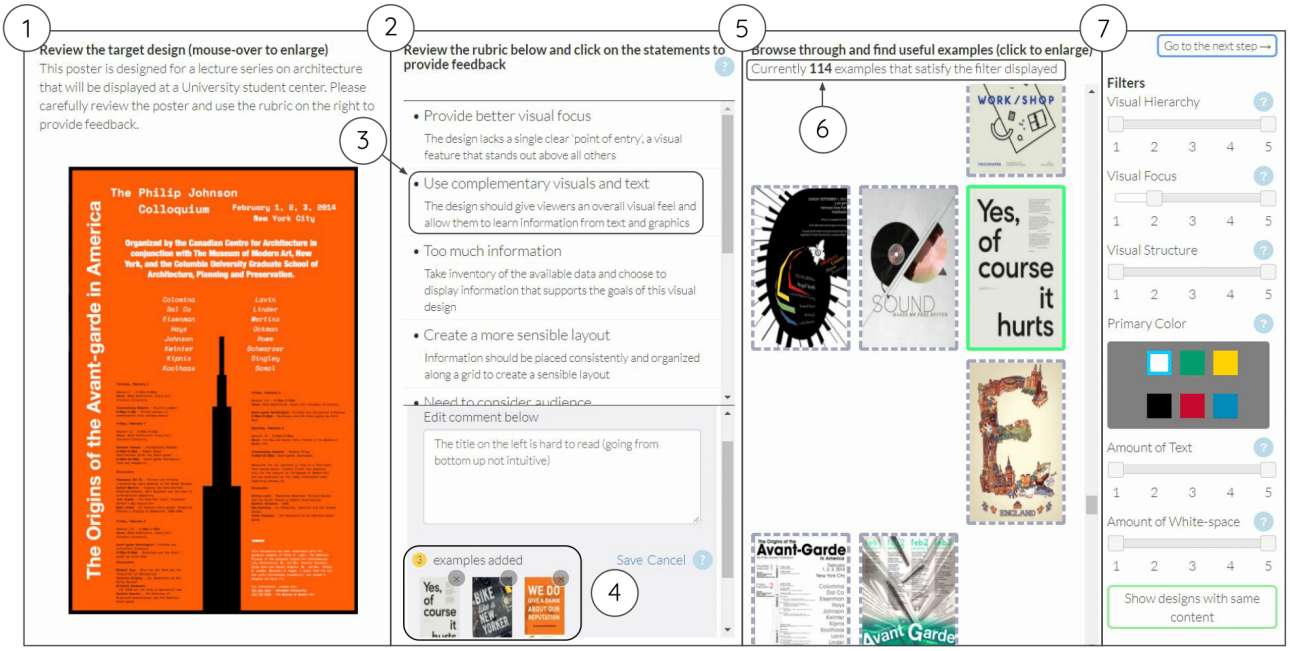

Figure 9. ***Paragon*****, an online gallery with filters for providing design feedback with examples. (1) shows the target design and explanation. (2) Feedback providers review the expert-created rubric and type their feedback. (3) Textbox for feedback appears when a rubric item is clicked. (4) Examples are shown at the bottom of the feedback. (5) shows the gallery of examples (in Examples and Examples+Metadata). (6) The number of displayed examples. (7) shows the filter widget; feedback providers can apply multiple filters at once (in Examples+Metadata). In Control, only panel 1 and 2 were displayed. In Examples, panel 7 was hidden.**

Each pair is from two different feedback providers with similar expertise.

Among all possible pairs created from 386 pieces of feedback, 2,536 and 1,700 pairs satisfied the conditions for the target design one and two, respectively. From this, we selected 156 (82 and 74 pairs for target design one and two, respectively) for grading. Among them, 15 for each target design were graded by all experts to compute the grading reliability. The reliability was measured by Intraclass Correlation Coefficients (ICCs) based on these grades. Among the 30 pairs graded together, a two-way mixed model [42] of ICC was 0.56 in specific, 0.50 in actionable, and 0.58 in novel dimensions, showing a moderate level of reliability [22]. Therefore, the remaining 126 pairs were distributed between four experts.

STUDY 2: RESULTS

Judges rated feedback with examples as higher quality

The difference between example-supplemented and text-only feedback was significant ($\chi^2(2) \geq 57.7, p < 3.0 \times 10^{-13}$) in all three dimensions. For example, in the Specific dimension, feedback supplemented with examples scored higher in 69 pairs. Target designs did not affect quality scores significantly. Therefore, we report only the aggregated result (table 2).

Feedback was longer when supplemented with examples

The mean character length of feedback was 40.1 (SD = 31.8) in Control and 191.7 (SD = 113.4) in the aggregated two example conditions. The condition of experiment had a significant effect on the length of feedback ($F(2,121) = 36.12, p < .001$). Comparative quality scores differed significantly with the difference in the length of feedback ($F(1,154) \geq 26.43, p < .001$

Dimension	Better w/ ex	Better w/o ex	Tie	%
Specific	69	33	5	64
Actionable	69	36	2	64
Novel	86	17	4	80

Table 2. Feedback with examples was judged as higher quality 64%, 64%, and 80% times in three dimensions among the 107 pairs that compared feedback with examples with text-only feedback. Better w/ example, Better w/o example, and Tie columns denote the feedback supplemented with examples being judged higher, lower, and equal in grades, respectively. 49 pairs had either both pieces of feedback supplemented by examples or no examples at all.

in three dimensions). An example pair of feedback (fig. 10, top) that scored 4 (i.e. "feedback with examples was greatly better") from all experts shows this tendency: feedback 1 (Control) was 27-characters-long while feedback 2 (Examples+Metadata) was 370-characters-long. When asked "how did examples affect the quality of feedback?", experts mentioned that examples helped with understanding issues by increasing the descriptiveness of explanation and providing reference points:

> *"when people are able to point to specific examples of things that didn't work and explain why they felt that way, it was easier for me to understand what needed to be changed. The more descriptive it was, the better I could envision the steps that would have to be taken"* – E3

In addition, feedback supplemented with examples had "better justification which led to more informative feedback" (E4) and brought in "more insight and inspiration" (E1) as well as "different perspectives of each facet (of design)" (E4).

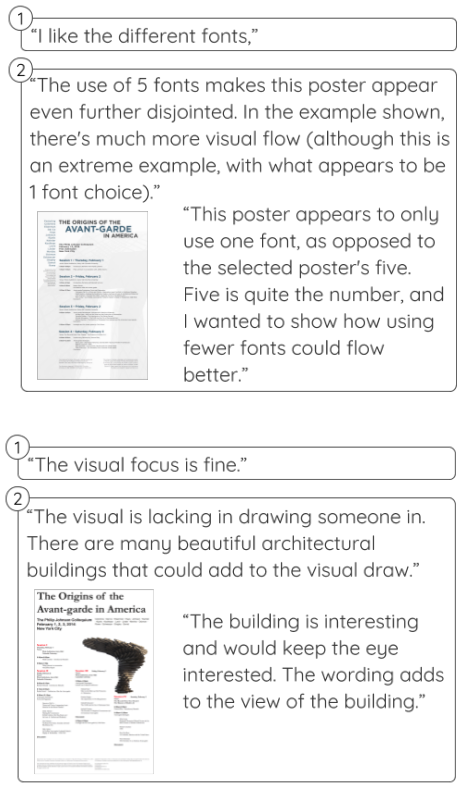

Figure 10. Two pairs judged "feedback with examples is greatly better" (score 4) in the Specific dimension. In each pair, feedback 1 (top) is from Control and 2 (bottom) is from Examples+Metadata. The top pair is for a rubric item "Thoughtfully choose the typeface and colors." The bottom pair is for "Provide better visual focus."

More time required to add examples to feedback

Providing participants an extra gallery and asking them to select examples to supplement their feedback likely increased the amount of time spent to complete the task. With application of Welch correction to account for difference in variances (Levene's test showed the variance of completion time is larger in example conditions. Target designs did not affect these variances), the result of one-way ANOVA shows a significant effect of experiment condition on the amount of time spent to provide feedback ($F(2,74.71) = 17.7, p < 1.00 \times 10^{-6}$). Participants spent on average 15.3 min (SD = 10.0) in the Examples+Metadata condition, followed by 13.4 min (SD = 8.9) in Examples and 6.2 min (SD = 5.6) in Control.

However, online participants considered the extra amount of time required in example conditions still acceptable. On average, their agreement to "I think the compensation was fair for the task" was 5.31 in Control (SD = 1.26), 4.7 (SD = 1.76) in Examples, and 4.81 in Examples+Metadata (SD = 1.74) on a 7-point Likert-scale, without any significant pairwise difference between conditions (Wilcoxon rank-sum test, $W \geq 809; p > .16$).

Participants used filters to hone into relevant examples

Some participants thought that the number of examples provided was overwhelming: *"200+ examples for us to choose from is overkill"* (P62). With such an abudance of examples, participants seeked to narrow down options using filters based on colors ("*It was useful for finding posters of certain colors*" – P100), content ("*It was most helptful to look at examples with the same content*" – P114), amount of text ("*(they were) useful especially when looking for images that had a lot of text to compare the main image to*" – P19), and layout ("*they were useful for making selections based on spacing and layout*" – P48). In addition, they used the filters to assist browsing (*"I think they (filters) are incredibly useful. They led me to one of the examples I chose, but they're also just helpful to make browsing through a huge number of images"* – P67). Some of this appreciation was complemented by comments made by participants in the Examples condition (who did not have access to filters). To an open-ended question "Are there any features that you wish you had in finding useful examples?", participants mentioned an ability to search by color (P8, 43, 68, 88, 116), font (P116), keyword (P14, 62), and the subject matter (P108).

However, no significant effects observed with filters

There was no significant difference between the Examples and the Examples+Metadata condition in the level of agreement to "It was easy to choose useful examples" (Wilcoxon rank-sum test, $W = 538, p = .60$); among the 40 subjects in Examples and 29 in Examples+Metadata who used at least one filter, the mean was 4.78 (SD = 1.70) in Examples and 5.07 (SD = 1.44) in Examples+Metadata. In addition, there was no significant difference between the conditions on the amount of time spent, the number of supplementary examples, or the length of feedback. Moreover, there was no significant difference between the conditions on their comparative quality scores (two-sided one-sample Wilcox signed-rank test comparing the sample median score to the mid-point, 2.5, $p > .27$. The distributions of scores departed from normality significantly $p < .0005$; Shapiro-Wilk test); among the 43 graded pairs that had feedback from both Examples and Examples+Metadata conditions, the mean score was 2.55 (SD = .78) in Specific, 2.56 (SD = .77) in Actionable, and 2.62 (SD = .76) in Novel dimensions where 1 = "feedback from Examples is greatly better" and 4 = "feedback from Examples+Metadata is greatly better."

STUDY 2: DISCUSSION

Study Two shows that the quality of feedback supplemented with examples was considered more specific, actionable, and novel by independent judges. Expert judges thought the descriptiveness of design issues and steps for improvement were enhanced by supplementary examples. However, online participants spent considerably more time when using examples to provide feedback and in turn, provided longer feedback. This did not mean participants were less willing to supplement feedback with examples; participants considered compensation for the task equally fair in experiment conditions.

When provided with a gallery of 287 design examples, participants used Paragon's filters enabled by crowdsourced metadata to narrow down options and browse for examples that fit best for feedback. Participants thought filters were useful for matching low-level design attributes among the examples (e.g. colors, amount of text, etc.), or to just assist their browsing.

However, the use of filters did not result in any significant effect on the perceived quality of produced feedback, subjective ratings on the ease of choosing useful examples, the amount of time spent to provide feedback, the length, or the number of supplementary examples. Some of this can be explained by additional cognitive load that the filter interface introduced. For example, some participants thought there was a learning curve to grasp how the filters worked, especially when they did not have previous design experience. This problem is potentially aggravated by the use of crowdsourced metadata. Although metadata for each design example was aggregated from multiple crowd workers and each worker was randomly assigned with at least 31 designs for consistency, the result may differ from expert-created metadata.

Some participants were concerned that using filters will lead to missing serendipitous discoveries. The effectiveness of such mechanisms, therefore, should consider not only their understandability and intuitiveness, but also the potential for serendipitous discovery. One way to increase the understandability and intuitiveness may be transparently showing how the gallery updates to respond to users' interest. Lee et al. have explored this idea in an adaptively updating gallery of examples [28]. One alternative approach to mitigating these issues may be using a bottom-up, example-oriented search for concept learning [10]. Leveraging online participants' ability to recognize and categorize interesting designs can remove the interpretive gap imposed by forcing them to follow the predetermined categories. However, having participants first build their own categorization may require more time in learning and completing the task. Drawing from child development in relational similarity [23] and a pilot with expert designers, we have created an interface with four low-level design-attribute filters (primary color, content-alignment, amount of text and white-space) and three high-level interpretive filters (the effectiveness of visual hierarchy, focus, and structure). Anecdotally, novice participants used low-level attributes more intuitively but struggled with interpretive filters. However, the seven filters we created are by no means a holistic representation of all dimensions for posters. Structuring workflows to crowdsource potentially more detailed and useful categories [4] as an alternative to experts' categorization would be valuable.

LIMITATION AND FUTURE WORK

One limitation of our research is that we have not examined the effect of feedback on final artifacts. Though it was shown by Luther et al. that higher-quality feedback from online crowds can help designers and lead to better final design artifacts [29], further investigation of how example-supplemented feedback is integrated in the design process will be fruitful. Along with other studies on reception and implementation of feedback (see, for example, [31, 33]), a particularly interesting issue for investigation is the design-fixation effect [36] and whether supplementing written feedback with examples causes recipients to emulate the exemplar. This could have positive effects on quality but would decrease the novelty of recipients' ideas. In addition, a closer look of the role and reception of examples in relation to content-similar, peer-created designs will have many implications on feedback processes in design classes. Though we demonstrated the efficacy of Web-collected examples in design feedback, what exactly are the 'good' and 'bad' examples, especially with relation to design fixation? Is artificially exposing students to distant examples beneficial?

While examples led to higher quality feedback in Study Two, participants also spent more time. This is likely attributed to the structure of the feedback process. In order to disambiguate the effect of rubrics and supplementary examples, we designed a two-stage process that requests feedback providers to divide their choice of examples and explanation specific to them in separate stages. The effect of such scaffolded reflection in the second stage of the process may be particularly helpful for novice online feedback providers. Building upon prior research [51], future work will investigate the role of reflection in this context and will focus on teasing apart the factors of time and structure.

Example-centric approaches have many future applications. Given the use of collages in design classes (e.g. inspiration or mood boards), curating example-centric feedback for effective presentation and reception will be useful. On a similar note, Kerne et al. have investigated strategies of free-form Web curation that stimulate students' creative engagement [20]. Another interesting investigation will be the reusability of example-supplemented feedback; because examples and the target design forms dyadic associations through feedback, a chain of feedback may be mined automatically from the network of how designs are associated with each other. Moreover, investigating not only the topics in feedback but also its temporal structure may give insights on how examples are used during different design stages and how might we reuse them in subsequent iterations (e.g. Related to this question, Kim et al. have explored how to create an online feedback-exchange platform specific to design works-in-progress [21]).

CONCLUSION

In this paper, we explored the efficacy, role, and benefit of using examples in an online design feedback process. Two randomized Web experiments showed that online novice participants were capable of using examples and preferred to do so in providing feedback. Feedback supplemented with examples were considered more specific, actionable, and novel by experts and they also tended to be longer than text-only feedback. This required more time to complete, however this additional workload did not seem detrimental, as measured by subjective scores of compensation fairness. Finally, we have built an online gallery interface with filters using crowdsourced metadata to aid the process of finding useful design examples for feedback providers. While some participants responded favorably, others expressed frustration, surfacing the challenges in striking a balance between the additional cognitive load and the usefulness of such a mechanism.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation award 1122206 and 1122320. We thank Dr. Scott Klemmer and The Design Lab UC San Diego for their generous feedback on the draft. We also thank online and offline participants of our study.

REFERENCES

1. Paul André, m.c. schraefel, Jaime Teevan, and Susan T. Dumais. 2009. Discovery is Never by Chance: Designing for (Un)Serendipity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition (C&C '09)*. ACM, New York, NY, USA, 305–314. DOI: http://dx.doi.org/10.1145/1640233.1640279
2. Margaret A. Boden. 1991. *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc., New York, NY, USA.
3. Joel Brandt, Philip J. Guo, Joel Lewenstein, Mira Dontcheva, and Scott R. Klemmer. 2009. Two Studies of Opportunistic Programming: Interleaving Web Foraging, Learning, and Writing Code. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1589–1598. DOI: http://dx.doi.org/10.1145/1518701.1518944
4. Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1999–2008. DOI: http://dx.doi.org/10.1145/2470654.2466265
5. ConceptArt. n.d. ConceptArt.org. http://www.conceptArt.org. (n.d.). [Accessed: 12-June-2017].
6. Deanna P. Dannels and Kelly Norris Martin. 2008. Critiquing Critiques: A Genre Analysis of Feedback Across Novice to Expert Design Studios. *Journal of Business and Technical Communication* 22, 2 (2008), 135–159. DOI: http://dx.doi.org/10.1177/1050651907311923
7. Steven P. Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel L. Schwartz, and Scott R. Klemmer. 2012a. *Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results*. Springer Berlin Heidelberg, Berlin, Heidelberg, 47–70. DOI:http://dx.doi.org/10.1007/978-3-642-31991-4_4
8. Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2012b. *Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-efficacy*. Springer Berlin Heidelberg, Berlin, Heidelberg, 127–153. DOI:http://dx.doi.org/10.1007/978-3-642-21643-5_8
9. Martin Wattenberg Fernanda Viégas. 2015. Design and Redesign in Data Visualization. https://medium.com/@hint_fm/design-and-redesign-4ab77206cf9. (2015). [Accessed 13-September-2017].
10. James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 29–38. DOI: http://dx.doi.org/10.1145/1357054.1357061
11. Dedre Gentner and Julie Colhoun. 2010. *Analogical Processes in Human Thinking and Learning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 35–48. DOI: http://dx.doi.org/10.1007/978-3-642-03129-8_3
12. Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology* 95, 2 (2003), 393. DOI: http://dx.doi.org/10.1037/0022-0663.95.2.393
13. Mary L. Gick and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1 – 38. DOI: http://dx.doi.org/10.1016/0010-0285(83)90002-6
14. David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* 35, 12 (Dec. 1992), 61–70. DOI: http://dx.doi.org/10.1145/138859.138867
15. Gabriela Goldschmidt, Hagay Hochman, and Itay Dafni. 2010. The design studio "crit": Teacher-student communication. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 24, 3 (2010), 285–302. DOI: http://dx.doi.org/10.1017/S089006041000020X
16. Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 235–244. DOI: http://dx.doi.org/10.1145/2757226.2757249
17. B. Hartmann, S. Doorley, and S. R. Klemmer. 2008. Hacking, Mashing, Gluing: Understanding Opportunistic Design. *IEEE Pervasive Computing* 7, 3 (July 2008), 46–54. DOI:http://dx.doi.org/10.1109/MPRV.2008.54
18. Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting Inspired!: Understanding How and Why Examples Are Used in Creative Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 87–96. DOI: http://dx.doi.org/10.1145/1518701.1518717
19. Benjamin Mako Hill and Andrés Monroy-Hernández. 2013. The Remixing Dilemma: The Trade-Off Between Generativity and Originality. *American Behavioral Scientist* 57, 5 (2013), 643–663. DOI: http://dx.doi.org/10.1177/0002764212469359
20. Andruid Kerne, Nic Lupfer, Rhema Linder, Yin Qu, Alyssa Valdez, Ajit Jain, Kade Keith, Matthew Carrasco, Jorge Vanegas, and Andrew Billingsley. 2017. Strategies of Free-Form Web Curation: Processes of Creative Engagement with Prior Work. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. ACM, New York, NY, USA, 380–392. DOI: http://dx.doi.org/10.1145/3059454.3059471

21. Joy Kim, Maneesh Agrawala, and Michael S. Bernstein. 2017. 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 (CSCW '17)*. ACM, New York, NY, USA, 246–258. DOI: http://dx.doi.org/10.1145/2998181.2998195

22. Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155 – 163. DOI: http://dx.doi.org/10.1016/j.jcm.2016.02.012

23. Laura Kotovsky and Dedre Gentner. 1996. Comparison and Categorization in the Development of Relational Similarity. *Child Development* 67, 6 (1996), 2797–2822. DOI: http://dx.doi.org/10.1111/j.1467-8624.1996.tb01889.x

24. Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M. Gerber, Brian P. Bailey, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4627–4639. DOI: http://dx.doi.org/10.1145/3025453.3025883

25. Chinmay Kulkarni, Steven P. Dow, and Scott R. Klemmer. 2014. *Early and Repeated Exposure to Examples Improves Creative Work*. Springer International Publishing, Cham, 49–62. DOI: http://dx.doi.org/10.1007/978-3-319-01303-9_4

26. Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*. ACM, New York, NY, USA, 75–84. DOI: http://dx.doi.org/10.1145/2724660.2724670

27. Ranjitha Kumar, Jerry O. Talton, Salman Ahmad, and Scott R. Klemmer. 2011. Bricolage: Example-based Retargeting for Web Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2197–2206. DOI: http://dx.doi.org/10.1145/1978942.1979262

28. Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R. Klemmer. 2010. Designing with Interactive Example Galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2257–2266. DOI: http://dx.doi.org/10.1145/1753326.1753667

29. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 473–485. DOI: http://dx.doi.org/10.1145/2675133.2675283

30. Jennifer Marlow and Laura Dabbish. 2014. 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 (CSCW '14)*. ACM, New York, NY, USA, 922–933. DOI: http://dx.doi.org/10.1145/2531602.2531651

31. Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science* 37, 4 (01 Jul 2009), 375–401. DOI: http://dx.doi.org/10.1007/s11251-008-9053-x

32. Mark W. Newman and James A. Landay. 2000. Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '00)*. ACM, New York, NY, USA, 263–274. DOI: http://dx.doi.org/10.1145/347642.347758

33. Thi Thao Duyen T. Nguyen, Thomas Garncarz, Felicia Ng, Laura A. Dabbish, and Steven P. Dow. 2017. Fruitful Feedback: Positive Affective Language and Source Anonymity Improve Critique Reception and Work Outcomes. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1024–1034. DOI: http://dx.doi.org/10.1145/2998181.2998319

34. Barbara Oakley, Richard M Felder, Rebecca Brent, and Imad Elhajj. 2004. Turning student groups into effective teams. *Journal of student centered learning* 2, 1 (2004), 9–34.

35. Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.

36. A.T. Purcell and J.S. Gero. 1992. Effects of examples on the results of a design activity. *Knowledge-Based Systems* 5, 1 (1992), 82 – 91. DOI: http://dx.doi.org/10.1016/0950-7051(92)90026-C Artificial Intelligence in Design Conference 1991 Special Issue.

37. Howard Risatti. 1987. Art Criticism in Discipline-Based Art Education. *Journal of Aesthetic Education* 21, 2 (1987), 217–225. http://www.jstor.org/stable/3332751

38. Daniel Ritchie, Ankita Arvind Kejriwal, and Scott R. Klemmer. 2011. D.Tour: Style-based Exploration of Design Example Galleries. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 165–174. DOI: http://dx.doi.org/10.1145/2047196.2047216

39. David A. Robb, Stefano Padilla, Britta Kalkreuter, and Mike J. Chantler. 2015. Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1355–1364. DOI: http://dx.doi.org/10.1145/2702123.2702470

40. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (01 Jun 1989), 119–144. DOI: http://dx.doi.org/10.1007/BF00117714

41. Nadine B. Sarter and David D. Woods. 1991. Situation Awareness: A Critical But Ill-Defined Phenomenon. *The International Journal of Aviation Psychology* 1, 1 (1991), 45–57. DOI: http://dx.doi.org/10.1207/s15327108ijap0101_4

42. Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420. DOI: http://dx.doi.org/10.1037/0033-2909.86.2.420

43. Pao Siangliulue, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 83–92. DOI: http://dx.doi.org/10.1145/2757226.2757230

44. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1243–1252. DOI: http://dx.doi.org/10.1145/1124772.1124960

45. Edward R Tufte, Susan R McKay, Wolfgang Christian, and James R Matey. 1998. Visual explanations: images and quantities, evidence and narrative. *Computers in Physics* 12, 2 (1998), 146–148. DOI: http://dx.doi.org/10.1063/1.168637

46. Anne Venables and Raymond Summit. 2003. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* 40, 3 (2003), 281–290. DOI: http://dx.doi.org/10.1080/1470329032000103816

47. Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 319–326. DOI: http://dx.doi.org/10.1145/985692.985733

48. Anbang Xu and Brian Bailey. 2012. 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 (CSCW '12)*. ACM, New York, NY, USA, 295–304. DOI: http://dx.doi.org/10.1145/2145204.2145252

49. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1433–1444. DOI: http://dx.doi.org/10.1145/2531602.2531604

50. Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 401–408. DOI: http://dx.doi.org/10.1145/642611.642681

51. Yu-Chun Grace Yen, Steven P. Dow, Elizabeth Gerber, and Brian P. Bailey. 2017. 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 (C&C '17)*. ACM, New York, NY, USA, 158–170. DOI: http://dx.doi.org/10.1145/3059454.3059468

52. Lixiu Yu and Jeffrey V. Nickerson. 2011. Cooks or Cobblers?: Crowd Creativity Through Combination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1393–1402. DOI: http://dx.doi.org/10.1145/1978942.1979147

53. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. 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 (CSCW '16)*. ACM, New York, NY, USA, 1005–1017. DOI: http://dx.doi.org/10.1145/2818048.2819953

54. ZURB. 2015. Forrst. http://zurb.com/forrst. (2015). [Accessed: 15-April-2017].