Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback

Fritz Lekschas
lekschas@seas.harvard.edu
Harvard School of Engineering and Applied Sciences
Cambridge, MA, USA

Spyridon Ampanavos
sampanavos@gsd.harvard.edu
Harvard Graduate School of Design
Cambridge, MA, USA

Pao Siangliulue
pao@b12.io
B12
New York City, NY, USA

Hanspeter Pfister
pfister@seas.harvard.edu
Harvard School of Engineering and Applied Sciences
Cambridge, MA, USA

Krzysztof Z. Gajos
kgajos@eecs.harvard.edu
Harvard School of Engineering and Applied Sciences
Cambridge, MA, USA

ABSTRACT

Crowdsourced design feedback systems are emerging resources for getting large amounts of feedback in a short period of time. Traditionally, the feedback comes in the form of a declarative statement, which often contains positive or negative sentiment. Prior research has shown that overly negative or positive sentiment can strongly influence the perceived usefulness and acceptance of feedback and, subsequently, lead to ineffective design revisions. To enhance the effectiveness of crowdsourced design feedback, we investigate a new approach for mitigating the effects of negative or positive feedback by combining open-ended and thought-provoking questions with declarative feedback statements. We conducted two user studies to assess the effects of question-based feedback on the sentiment and quality of design revisions in the context of graphic design. We found that crowdsourced question-based feedback contains more neutral sentiment than statement-based feedback. Moreover, we provide evidence that presenting feedback as questions followed by statements leads to better design revisions than question- or statement-based feedback alone.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

crowdsourced design feedback, feedback framing, sentiment, questioning

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445507

ACM Reference Format:
Fritz Lekschas, Spyridon Ampanavos, Pao Siangliulue, Hanspeter Pfister, and Krzysztof Z. Gajos. 2021. Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445507

1 INTRODUCTION

Feedback is a central part of learning and achievement that can help evaluate one's work, uncover problems, and promote new ideas for improvement. Yet, its effectiveness greatly varies by type and how it is framed, and its impact can be either positive or negative [20]. In graphic design, feedback is a vital part of the iterative design process and is typically solicited in critique sessions. However, these sessions are time and resource intensive. Moreover, feedback from alternative sources like peers and online communities can be scarce [30, 31, 51], biased [44, 51], and superficial [46, 50]. Crowdsourced online feedback is an emerging mechanism to gather large amounts of feedback quickly [18, 29, 53]. When structured appropriately, crowdsourced feedback can be as effective as expert feedback [55] and help designers produce more and better design revisions than they could have done otherwise [30, 51, 52].

For crowdsourced feedback to be effective, it needs to foster productive reflection on the design to generate useful ideas for design revisions. Furthermore, the feedback needs to be acceptable to the designer, or else they will ignore it. However, this is challenging because there is a tension between the productive value of feedback and acceptability, which is related to the feedback's perceived sentiment. For instance, Crain et al. [12] found that feedback with positive sentiment, which we will refer to as positive feedback, is typically preferred by content creators. However, positive feedback is less likely to lead to improvements through iteration. On the other hand, in their study, feedback with negative sentiment encouraged more design iterations but tended to have lower acceptance. In the worst case, feedback with negative sentiment, which we will refer to as negative feedback, influences the recipient's affective state [1, 49] and can reduce their overall task performance [5].

To improve the effectiveness of crowdsourced feedback on design revisions, we contribute a novel approach of enhancing traditional statement-based feedback with open-ended and thought-provoking

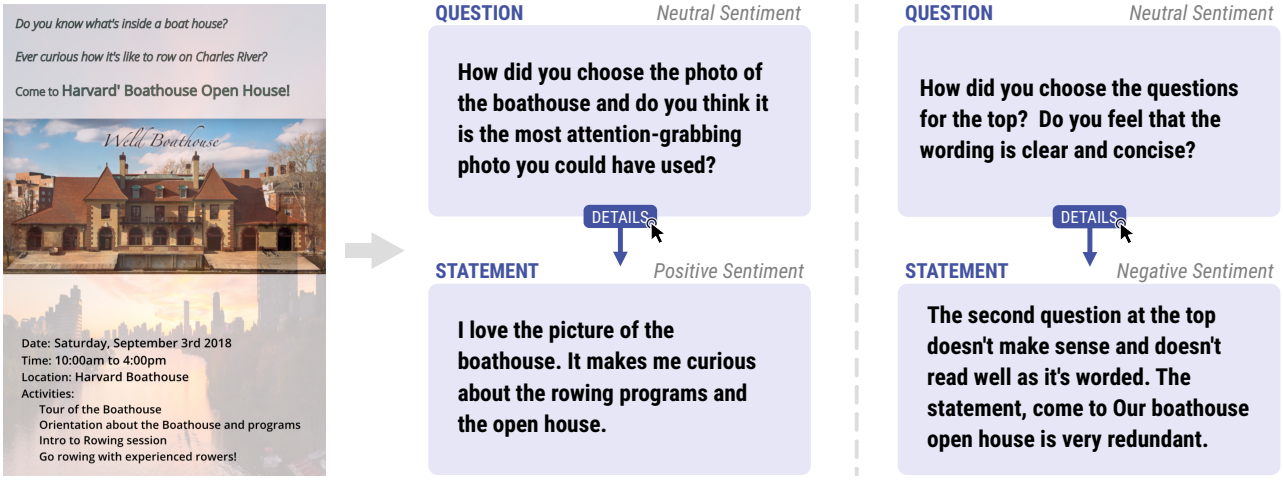

Figure 1: Enhanced Design Feedback: Two example feedback items for a flyer from the first user study (Section 4). Each feedback item consists of an open-ended question followed by a traditional statement. Although the related questions and statements target the same aspects of the flyer design, the questions carry more neutral sentiment than the statements.

questions (Figure 1). We hypothesized that presenting feedback in the form of a question followed by a statement would result in higher-quality design revisions compared to statement-based or question-based feedback alone. Building on prior work from several fields, our rationale for this hypothesis is twofold. First, we hypothesized that feedback in the form of open-ended questions carries less sentiment than statements and, subsequently, improves the acceptance of the feedback. Second, we hypothesized that the preceding open-ended question promotes productive reflection even if the statement-based feedback is superficial or unacceptable to the designer.

In design, reflection is fundamental in evaluating the current state of one's work relative to its goals and for generating ideas for improvements [39]. It is suggested that combining feedback with reflection is a superior format [4] compared to feedback alone. For instance, feedback that incorporates a reflective task can lead to more extensive revisions and increased quality [54] compared to traditional feedback. An effective way to promote reflection is facilitative questioning. For example, in teaching, questioning is known as an effective technique to trigger reflection and critical thinking among students [6, 43]. However, questioning should not be the only type of feedback as it can otherwise irritate students [3]. Besides reflection, questions could balance the acceptance of feedback statements, assuming they contain neutral sentiment. For instance, ordering feedback from positive to negative has been shown to lead to a more balanced perception of negative feedback by improving the recipients' happiness and excitement [48].

We conducted two online user studies in the context of graphic design to study the effects of enhancing statement-based with question-based feedback. In the first study, we investigated if feedback in the form of open-ended and thought-provoking questions can be crowdsourced and if these questions contain more neutral sentiment compared to corresponding feedback statements. The results show that 85% of the questions created by the crowd workers are open-ended and thought-provoking. We also found that the questions derived from negative or positive statements contained significantly more neutral sentiment than the corresponding statements, as exemplified in Figure 1. In the second study, we examined the effectiveness of feedback enhanced with open-ended questions on the quality of design revisions. We recruited 36 non-professional designers to design a flyer and revise it based on crowdsourced feedback. To test our hypothesis, we assessed three ways of presenting the feedback: statements only, questions only, and questions followed by statements. We employed an external jury of expert designers to rate the flyers' design quality for comparison. We found that participants who were shown questions followed by statements improved their designs to a significantly greater degree than participants who saw either statements or questions alone.

We make two contributions to the area of crowdsourced design feedback. First, we introduce the first method for framing crowdsourced design feedback as questions and combining them with traditional feedback statements. Second, we provide empirical evidence that presenting crowdsourced feedback in the form of open-ended questions followed by statements improves the quality of design revisions compared to presenting feedback as either statements or questions alone. Combining statement-based feedback with open-ended questions is complementary to other strategies for enhancing the effectiveness of design feedback. Therefore, our approach can easily be integrated into existing crowdsourced design feedback systems to increase the overall productive value of the feedback for design revisions.

2 RELATED WORK

2.1 Background

Within the inherently iterative design process, feedback is essential to evaluate the design's current state and generate revision ideas [20, 36]. Design studios are a fundamental element in design education, where students receive feedback in various types of

critique sessions [40]. These critique sessions consist of a work presentation by the student followed by an individual critique from the teacher (i.e., "desk crit"), multi-layered critique by a jury, or open feedback from other students [45]. Ideally these sessions result in a dialogue for finding a common ground between one's own design intentions and the received feedback. In the professional practice, designers are seeking such detailed feedback from peers. Overall, design critiques provide in-depth analyses and foster a deep understanding of the designer's work [10, 13]. However, while providing rich feedback, critiques can be infrequent, time-consuming, and resource-intensive. Therefore, designers may require additional feedback in preparation for the more structured critique sessions. Peers and online communities can provide such additional feedback but it can be limited in quantity [30, 31, 51], biased [44, 51], and superficial [46, 50]. Crowdsourcing is an approach to overcome these limitations [18, 29, 53] and provide almost expert-quality feedback when elicited and structured effectively [55].

2.2 Sentiment and Valence

Prior research on crowdsourced feedback systems found that the sentiment of feedback impacts its perceived usefulness. For example, Yuan et al. [55] found that "positively written and emotional critiques received higher average ratings". Their findings provide evidence that valence and arousal are positively correlated with designers' ratings of feedback. Similarly, Nguyen et al. [33] studied feedback on writing tasks and found that positive tone in critical feedback leads to better work quality overall. Krause et al. [28] systematically investigated the perceived usefulness of feedback along various dimensions such as length, specificity, or complexity. They found that the perceived usefulness peaks for feedback with neutral to very mildly negative sentiment. Wu et al. [48] build upon these findings and studied the effects of presenting feedback with varying sentiments in different orders. They present empirical evidence that showing negative feedback at the end improved the feedback's perception.

However, in contrast to the perceived usefulness, Crain et al. [12] studied the long-term effects of different types of feedback on design iterations in a large meta-study on feedback collected from Reddit. They found that longer and less positive feedback is predictive of a higher number of design iterations. Although the study could only take publicly shared iterations into account, it highlights a disparity between the perceived usefulness and the actual effectiveness of feedback with diverging sentiment.

Sargeant et al. [37] studied the impact of positive and negative feedback on the recipient. They found that negative feedback can evoke negative feelings, especially when the feedback disagrees with the recipient's self-perception. In this case, the recipient perceives the feedback to be addressed against themselves rather than the task at hand. Wu et al. [49] confirmed these findings and additionally showed that balancing the valence of feedback can mitigate the impact of negative feedback on its perceived usefulness.

We hypothesize that framing feedback as a question will alleviate sentiment. Subsequently, we hypothesize that showing feedback in the form of questions prior to the traditional statement-based feedback will increase the feedback's overall acceptability.

2.3 Reflection

The ultimate goal of feedback is to help improve the critiqued work. In order to achieve this goal, feedback needs to facilitate new productive ideas. Beyond direct feedback, reflection is another popular tool [39] in the design community to generate ideas for design revisions. See Baumer et al. [2] for a review on how reflection can be leveraged in the design process as a whole. In regard to feedback, Caroline Brandt [4] showed that feedback alone might not always be sufficient. She suggests that combining feedback with a reflection task is generally superior.

Yen et al. [54] confirmed this hypothesis by showing that reflection alone can be as beneficial as crowdsourced feedback. They implement a reflective activity where designers have to respond to three generic questions about their design. In their study, the combination of reflection and feedback led to the best design quality overall. Moreover, Sargeant et al. [38] found that facilitated reflection can alleviate the distress caused by negative feedback and enhance feedback acceptance.

In this work, we build upon these findings and hypothesize that feedback in the form of questions will act as a lightweight reflective activity that promotes useful ideas for design revisions. Moreover, we extend previous reflection approaches by preceding a negative feedback statement with an open-ended question related to the same aspect of the design to help designers to better cope with potential distress caused by the negative feedback.

2.4 Facilitative Questioning

For questions to be effective, they need to facilitate reflection and promote critical thinking. For instance, in evaluating writing, Knoblauch and Brannon [27] have established an approach called "Facilitative Response", which argues that the reviewer should adopt a "facilitative posture". Instead of directly telling the writer what to do, the reviewer should raise open-ended questions to encourage the writer to think about their ideas and expressions more fully. Facilitative responses do not need to come in the form of questions, but studies have found questions to be an effective implementation.

For example, Carnine et al. [6] found positive effects for facilitative questioning in combination with feedback in teaching children. Berghmans et al. [3] studied the benefits of facilitative questioning against direct teaching approaches for medical students. They found that facilitative questioning is beneficial for students with less expertise. Interestingly, they also discovered that questioning alone is not perceived well as students demand information after facilitative questions were raised.

In general, questioning has been studied as a tool for teaching. For example, Alison King developed a technique called "reciprocal questioning" [24, 25] in which she provides evidence that thought-provoking questions lead to a deep discussion about topics and encourage critical thinking [26]. Ciardiello et al. [8] discuss how to identify and generate divergent questions to promote literacy. Chambers et al. [7] compared questioning as a teaching tool for swimmers and found that deliberately delaying extensive amounts of feedback and replacing it with insightful questions elicits better reflection and ultimately improves the swimmers' technique.

In our approach, we implement facilitative questioning as a tool to promote reflection and critical thinking.

2.5 Framing & Structuring Feedback

Irrespective of the feedback's sentiment and reflective nature, the way a system elicits and structures feedback from non-expert crowd workers can change the feedback's focus and quality. For example, Hicks et al. [21] investigate three different ways of framing feedback. They found that asking for numerical ratings of the design leads to more explanatory feedback of lower quality.

Sadler describes effective feedback to be specific (following a predefined concept), goal-oriented (comparing the work's current to a reference state), and actionable (promoting actions that close the performance gap) [36]. As elaborated by Connor and Irizarry, these three elements are equally necessary for design critiques [10]. They additionally argued that the critique's goal should be an analysis of the performance gap to drive effective design iterations. In the context of crowdsourcing, several studies [18, 23, 30, 32, 35, 51, 55] have evaluated the effects of structuring and scaffolding feedback and found that an appropriate structure elicits more diverse and higher quality feedback. For example, Voyant [51] prompts non-expert feedback providers to provide smaller feedback on various specific aspects of a design. In CrowdCrit [30], Luther et al. built upon these findings and further structured the feedback task into problem identification and explanation.

In our method, we utilize these findings by asking the feedback providers to focus on three different aspects of the design.

3 APPROACH AND HYPOTHESES

Previous research indicates a design tension (Section 2). Positive feedback is more acceptable to the recipient, but it is less likely to lead to substantial revisions compared to negative feedback. On the other hand, negative feedback can lead to substantial design improvements, but it is a source of discouragement and it is likely to be dismissed. This is particularly challenging in the context of crowdsourced design feedback systems, an otherwise promising source of feedback. How can we enhance crowdsourced design feedback to be acceptable and substantive to promote useful ideas for design revisions? And how can we elicit such feedback robustly from non-expert crowd workers?

Our approach is to structure feedback such that a potentially negative or positive statement is preceded by an open-ended question related to the same concern. For instance, in the context of designing an event flyer, "This image is not relevant to the event" might be preceded by "What made you choose this image?", or "How is this image related to the event?". To ensure that the question and statement relate to the same concern, the feedback provider is asked to first provide statement-based feedback and subsequently rephrase the statement into an open-ended and thought-provoking question. We consider a question to be open-ended when it requires an elaborating answer beyond "yes", "no", or simple facts. The goal of such a question is to promote critical thinking and reflection about a specific aspect of the critiqued work without carrying overly positive or negative sentiment.

In this context, our main hypothesis is the following:

H-Main: Feedback in the form of an open-ended question followed by a statement improves the overall quality of design revisions compared to statement-based or question-based feedback alone. Our reasoning is twofold. We hypothesize that the preceding question increases the acceptance of negative feedback and that asking a question will act as a lightweight reflective task, which can promote better design revision, as shown by Yen et al. [54]. However, we expect feedback consisting of questions alone to lead to less effective design revisions as it can irritate the feedback receiver [3].

To answer our main hypothesis, we pose the following supporting hypotheses on the effects of question-based feedback:

H-Support 1: Non-expert crowd workers can ask open-ended and thought-provoking questions. Given prior work on the effectiveness of structuring feedback acquisition (Section 2.5), in particular the work by Greenberg et al. [18], we hypothesize that providing a clear structure on how to provide feedback in combination with relevant example questions will teach the workers how to pose open-ended and thought-provoking questions, just like Alison King did with her students [24, 25].

H-Support 2: Feedback in the form of an open-ended question has more neutral sentiment than feedback addressing the same concern, but framed as a statement. Assuming that crowd workers are able to pose such questions, we hypothesize that open-ended questions carry more neutral sentiment than statements given the nature of open-ended questions.

H-Support 3: Preceding question-based feedback leads to more balanced acceptance of subsequent statement-based feedback compared to statement-based feedback alone. Assuming open-ended questions contain more neutral sentiment than statements and taking into account the improvement in perception of negative feedback when preceded by positive feedback [48], we hypothesize that presenting the question-based feedback first will cause the recipients to focus on the design rather than themselves and perceive subsequent statement-based feedback more neutrally compared to statement-based feedback alone.

4 STUDY 1: ELICITING OPEN-ENDED FEEDBACK QUESTIONS FROM CROWD WORKERS

In support of **H-Main**, we investigated if open-ended question-based feedback can be crowdsourced from non-experts (**H-Support 1**) and if such question-based feedback contains more neutral sentiment than statement-based feedback (**H-Support 2**). To this end, we asked online crowd workers to provide feedback for graphic designs in the form of statements and questions.

4.1 Experimental Design

In our approach (Section 3), we ask each feedback provider to rephrase their feedback statement into a question to ensure that the feedback addresses the same aspect of the design. However, the act of rephrasing might be a confounding factor that influences the sentiment and open-endedness. To control for this potential confounding factor, we conducted a within-subjects experiment with two factors: *framing* and *rephrasing*. Framing has two levels, which refer to posing feedback as either declaratory statements or open-ended questions. Rephrasing describes the strategy of eliciting statements-questions pairs and has the following two levels: rephrasing statements into questions (**S→Q**) or vice versa (**Q→S**).

4.2 Task

We presented each participant with four diverse designs of a flyer advertising a local event. We asked each participant to provide three written feedback items (addressing the *theme* of the design, the *layout* of the design, and a specific visual *element* in the flyer). For the first two flyers, the participants had to write a statement first and then rephrase it into a question (**S→Q**). For the other two flyers, the participant had to first write the question and then rephrase the question into a statement (**Q→S**). Following Greenberg et al. [18], we provided three diverse examples to promote creativity [41, 42] and encourage feedback that addresses a variety of aspects. Each example consisted of a statement and question.

4.3 Participants

We recruited 24 participants (16 male and 8 female) on Amazon Mechanical Turk (AMT) who were located in the US and spoke English natively. Only participants with an acceptance rate above 97% and more than 500 approved HITs were accepted. The majority of participants (16) were aged between 30–40. Three were between 20-30 years old. Another three were between 40-50 years old. And two were aged between 50–60. On average, the participants reported to be somewhat familiar with graphic design principles (M=3.17) and not very proficient in generating graphic designs (M=2.58). The results were reported on a 5-point Likert scale from "very unfamiliar" to "very familiar" and "very unproficient" to "very proficient" respectively. Participants were paid 5 USD for completing the task.

4.4 Procedure

We divided the participants into two groups, where the first group started with rephrasing statements into questions (**S→Q**) two times and then switched to **Q→S**. The second group started with **Q→S** and switched to **S→Q** after the first two flyers. Supplementary Figures S2–S4 show how the task was implemented. To avoid mistakes when the participants switched from **S→Q** to **Q→S** and vice versa, we added a dedicated step to inform about the upcoming switch in the rephrasing strategy. In total, each participant provided 12 feedback items: three feedback items for each of the four flyer designs. The order of the flyers was randomized.

4.5 Measurements

Open-endedness. We measured the rate of successfully-rephrased statements into open-ended and thought-provoking questions through coding. The first two authors of this paper coded all statements as being either successfully rephrased into open-ended and thought-provoking questions or not. We considered a question to be open-ended and thought-provoking if it required more than a yes/no answer or a statement of simple facts. Specifically, we used Alison King's [24–26] question stems (e.g., "How did you choose…", "What is the purpose of…", or "Why did you decide on…") as guidance and we assessed if the question targeted the rationale behind a design choice.

Prior to the analysis, feedback that did not target the actual design was removed. Such peripheral feedback questions typically focus on predefined requirements (e.g., "What made you name it Harvard Open Boathouse if it's technically not "open" to anyone except for Harvard students?") or facts about the photographic material (e.g., "Is this one of the actual boats that are currently being used by the crew?").

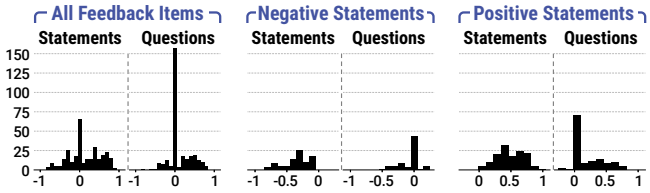

Figure 2: Feedback Polarity: Distribution of polarity scores (x-axes) across all feedback items (left), items related to negative statements (middle), and items related to positive statements (right). Questions have more neutral sentiment on average than the corresponding statements.

The authors initially coded all questions individually using separate Google Sheets with questions in randomized order. They achieved high agreement of Krippendorff's $\alpha = .81$ (calculated in *Python* using Grill's `krippendorff_alpha` method [19]). Subsequently, they collaboratively resolved conflicts to reach complete agreement. Most conflicts were due to two types of questions: questions that ask for a reason (e.g., "Is there some reason why you did not decide to go with a more blue color to kind of go along with boating?") and questions that ask for an alternative (e.g., "Does the text at the bottom contrast enough against the water? Is there another color that might work better?").

Sentiment. We analyzed the sentiment of every feedback statement and question using `VADER` [22]—an automated sentiment analysis tool. `VADER` provides a polarity score ranging from −1 to 1, where −1 refers to negative sentiment, 1 refers to positive sentiment. We consider scores between −0.05 and 0.05 as neutral sentiment.

4.6 Results

Ten out of 288 feedback questions (3.5%) were removed from the analysis as they did not pertain to the graphical design choices. Of the remaining 278 questions, 236 (84.9%) were found to be open-ended and thought-provoking.

The distribution of sentiment polarity scores for the statement- and question-based feedback items are shown in Figure 2. As confirmed by a Shapiro-Wilk test of normality, the polarity scores are not normally distributed (W=.92, p<.0001). Therefore, we conducted a Wilcoxon signed-rank test to compare the absolute polarity of statement-based and question-based feedback. We found that statement-based feedback had significantly higher absolute polarity (M=.33, SD=.27) than question-based feedback (M=.18, SD=.23; W=5703.5, p<.0001).

To better understand how the question and statement sentiment differed, we separately analyzed the polarity scores of statement-question pairs associated to statements with a polarity smaller than −.05 (i.e., negative statements), larger than .05 (i.e., positive statements), and polarity in [-.05, .05] (i.e., neutral statements). For negative statements (n=87), we found that statement-based feedback had significantly more negative polarity scores (M=-.34, SD=.20) than the related question-based feedback (M=0.07, SD=0.28; W=112, p<.0001). Similarly, for positive statements (n=128), statement-based

feedback had significantly higher polarity scores (M=.50, SD=.21) than the related question-based feedback (M=.17, SD=.27; W=643.0, p<.0001). For neutral statements (n=63), we did not find any significant difference in the scores for statement-based (M=.00, SD=.01) and question-based feedback (M=.04, SD=.21; W=89.5, p=.14).

To determine the influence of *rephrasing* (**S→Q** and **Q→S**), which might be a potential confounding factor (Section 4.1), we analyzed its impact on the questions' open-endedness and sentiment. Knowing the influence of rephrasing can also inform future practical uses of our method. A Cochran's Q test showed that there was no significant association between rephrasing and open-endedness of the questions (Q=8.92, p=.63).

Regarding the impact of rephrasing on the sentiment polarity scores, we were additionally interesting in testing for potential interactions effects between rephrasing and framing. To use a non-parametric factorial analysis, we first applied the Aligned Rank Transform [47] on the polarity scores. Using the aligned polarity scores, we conducted a repeated-measures analysis of variance (ANOVA) with framing and rephrasing as the two within-subjects factors. As expected, we observed a significant effect of framing on absolute polarity (F(1,552)=65.51, p<.0001) and no significant effect of rephrasing on the absolute polarity (F(1,552)=1.23, p=.27). We also did not find any significant interaction between framing and rephrasing (F(1,552)=1.46, p=.23).

We separately repeated the same analysis for question-statement pairs associated with negative and positive statements. For negative statements, we again find a significant effect for framing (F(1,170)=191.98, p<.0001) and no significant effect for rephrasing (F(1,170)=.53, p=.47). However, this time we found a significant interaction between framing and rephrasing (F(1,170)=5.41, p=.021). Investigating the simple main effects for **Q→S** and **S→Q** separately, we find that questions (M=.09, SD=.3) had a more neutral polarity score (**Q→S**: M=.09, SD=.3; **S→Q**: M=.05, SD=.27) than statements (**Q→S**: M=-.38, SD=.21; **S→Q**: M=-.31, SD=.19) in both cases (**Q→S**: F(1,86)=110.87, p<.0001; **S→Q**: F(1,84)=81.1, p<.0001). Similarly, for positive statements, we find a significant effect for framing (F(1,252)=115.92, p<.0001), no significant effect for rephrasing (F(1,252)=.96, p=.33), and a significant interaction between framing and rephrasing (F(1,252)=6.84, p=.01) We again investigated the simple main effects for **Q→S** and **S→Q** separately and found that questions had a more neutral polarity score (**Q→S**: M=.19, SD=.25; **S→Q**: M=.15, SD=.30) than statements (**Q→S**: M=.47, SD=.21; **S→Q**: M=.52, SD=.22) in both cases (**Q→S**: F(1,128)=48.95, p<.0001; **S→Q**: F(1,124)=68.67, p<.0001).

4.7 Summary and Discussion

The results of this study demonstrate that non-experts recruited online can produce open-ended questions with a high degree of success (84.9%), which supports **H-Support 1**. Our results also demonstrate that feedback phrased as questions has weaker polarity than equivalent feedback presented as declarative statements according to automated sentiment analysis. That is, questions related to negative feedback express more neutral sentiment than their corresponding statements, and questions related to positive feedback also express more neutral sentiment than statements expressing equivalent critique. These findings support **H-Support 2**. Finally, our results suggest that the order in which feedback is rephrased does not have a strong effect on the feedback's sentiment. While we see an interaction between framing and rephrasing, the simple main effects indicate that questions have significantly less sentiment compared to statement in both rephrasing orders.

One concern is the influence of the payment on the feedback. Prior research suggests that the principal effect of payment is the increased quantity of work: Unpaid crowds provide less feedback than paid workers [50, 51]. A factor that may be of greater relevance is anonymity, which can improve the feedback quality by avoiding peer pressure [31]. Thus, we assume that our results on the quality and sentiment of feedback will generalize to unpaid settings as long as the feedback is anonymous. However, more studies are necessary to verify this assumption.

5 STUDY 2: THE EFFECTS OF COMBINING STATEMENT- WITH QUESTION-BASED FEEDBACK

In the second user study, we examined our main hypothesis **H-Main** and the supporting hypothesis **H-Support 3** in the context of a graphic design task. The study consisted of two sessions. In the first session, participants designed an event flyer, for which we subsequently crowdsourced feedback. Based on this feedback, participants revised their initial design in the second session. Finally, an independent jury of design experts rated the improvements of the revised designs.

5.1 Experimental Design

We conducted a between-subjects experiment in which we compared the following three conditions: statement-based feedback only (**S**), question-based feedback only (**Q**), and question-based feedback followed by statement-based feedback (**Q+S**). While our main hypothesis (**H-Main**) is that the revision quality in **Q+S** will be higher than in **S**, we included **Q** to be able to determine whether the hypothesized improvement is due to the combination or framing of feedback. The participants were equally and randomly distributed across the three conditions.

5.2 Task

The participants were asked to design a flyer for a local sports event. The event, called "Harvard Open Boathouse" was a fictional open house day of a university-affiliated rowing club that invites university members to learn about the sport, facilities, and meet senior club members. We chose this fictional event to focus on a specific event type that is popular in the local area.

In the first session, participants designed their initial flyer, which they subsequently in the second session. Before revising their flyer design, the participants were presented with crowdsourced feedback (Figure 3), which we asked them to address in their revision. See Supplementary Figure S14 for a full example. During the feedback presentation, participants had to rate how much each statement or question made them think about their design in new ways. Since our goal was to capture the immediately-perceived *thought-provokingness* of each feedback item, the form fields disappeared after the corresponding feedback was rated. In the **Q+S** condition, the participants saw only the question-based feedback until they

rated the thought-provokingness, but a text label indicated that more information (i.e., the feedback statement) would appear after rating. In all conditions, participants were not allowed to proceed and upload their revised design until all feedback items had been rated. Inspired by Yen et al. [54], we wanted the participants to think about the question-based feedback explicitly to encourage reflection. Furthermore, in **Q+S**, we wanted to contrast the reported thought-provokingness against the final feedback ratings (Section 5.6) to assess whether preceding questions increase the perceived usefulness of the feedback.

5.3 Participants

Designers. We recruited 36 participants (8 male and 28 female) located around Harvard University (Cambridge, MA) using flyers and mailing lists. The majority of participants (21) were aged between 18–25 while the rest (15) were aged between 26–35. We targeted participants who were relatively inexperienced in graphic design, as prior research [3, 15] has shown that experienced designers have often built high confidence in their skill sets and rely primarily on their experience rather than feedback. In a pre-study questionnaire, most participants (25 out of 36) reported that they had never created a graphic design in a professional capacity. Per completion of both sessions, participants received a 35-USD gift card.

Feedback Providers. We recruited 187 participants on AMT to provide feedback on the flyer designs. As in the first study (Section 4), we only accepted US-based workers with an acceptance rate above 97% and more than 500 approved HITs. To prevent any potential learning effects and ensure an equal distribution of independent feedback providers per design, we used Unique Turker [34], which stopped feedback providers from completing the user study multiple times. For statement- (**S**) and question-only (**Q**) feedback, we paid 0.85 USD per task. For the combined feedback (**Q+S**), we paid 1.25 USD per task.

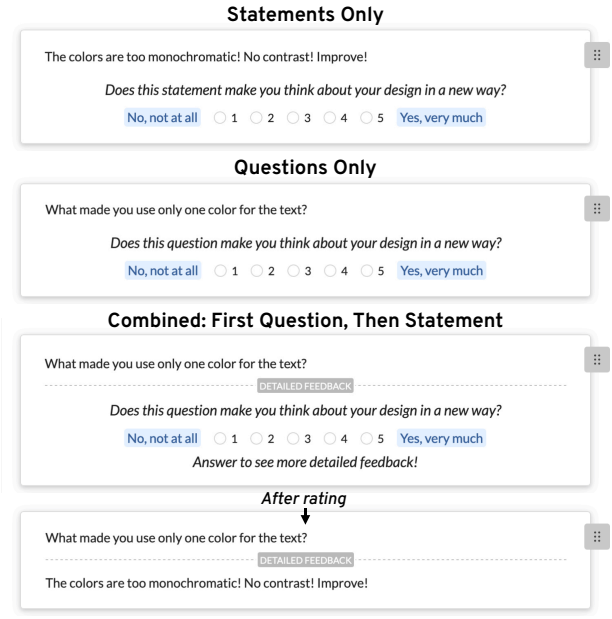

Figure 3: Feedback Presentation: In the combined condition (Q+S), the statement was only shown after the thought-provokingness was rated.

Judges. To evaluate and rate the improvement of the flyer designs, we recruited a jury of eight design experts (three male and five female). We considered someone to be a design expert if they hold an academic degree in a field related to graphic design, had at least two years of work experience as a professional designer or had taught at least one course related to graphic design. Three experts earned a doctor degree while the others held a master degree in architecture, UI/UX/HCI, or fine arts. Five judges were professors, two were graduate research assistants with teaching experience, and one was a professional designer. Each expert received a 50-USD gift card as compensation.

5.4 Main Study Procedure

We conducted the study online to allow participants to work on their designs anywhere and anytime. Our web application guided the participants through each step of the user study. See Supplementary Figures S5–S19 for a complete walkthrough. We split the experiment into two sessions to allow for enough time to collect feedback. Figure 4 shows an overview of the procedure.

The first session comprised the consent process, pre-study questionnaire, design brief, and the first design iteration. The participants were free to use their software of choice for designing the flyer. For participants who did not have access to any graphics software, we recommended Google Drawings [17] and Gravit Designer [11]. After each participant completed the first session, we acquired, filtered, and randomly selected crowdsourced feedback. In the second session, the participants were presented with the feedback, revised their initial design, rated the received feedback, and completed the post-study questionnaire. Each session took 45–60 minutes. We started measuring the time before presenting the instructions for designing and revising the flyer and showed a timer for convenience.

Finally, an independent jury of design experts rated the improvement of the design revisions and selected the three best designs. We randomized the order of the flyers for each jury member to avoid interaction effects between the flyer's position and rating. The participant with the highest average quality rating received an additional 100-USD gift card. We included the competition to increase the participants' motivation throughout the two sessions.

5.5 Acquisition and Selection of Crowdsourced Feedback

For each flyer design, we collected 15 feedback items from five unique crowd-workers (i.e., three feedback items per worker) using the **S→Q** feedback acquisition procedure from Section 4. Anticipating how the **S** and **Q** conditions might be implemented in practice, we asked the feedback providers to only give statement-based or question-based feedback, respectively. Hence, the rephrasing step was omitted in **S** and **Q**.

After collecting the feedback (Figure 5), the first two authors of this paper inspected each set of three feedback items to ensure a minimum level of quality. In 7 out of 180 cases, the crowd worker

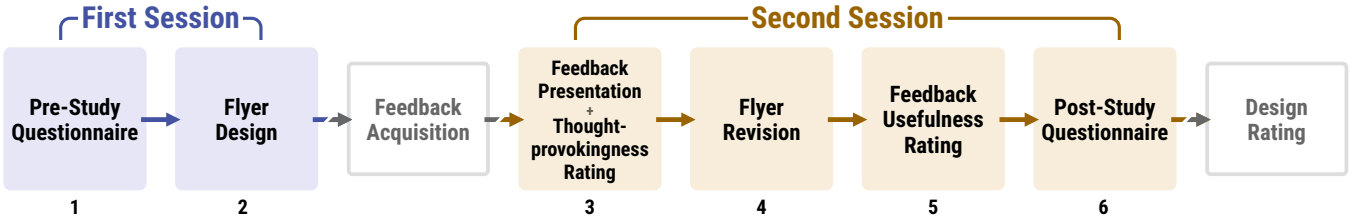

Figure 4: User Study Procedure: In the first session, the participants completed a pre-study questionnaire (1) and created an initial flyer design (2). Afterward, we crowdsourced feedback from AMT. (See Figure 1 for an example.) In the second session, the participants read the feedback (3), revised their design (4), rated the feedback (5), and completed a post-study questionnaire (4). Finally, a jury of design experts rated the improvement of the flyer designs (Figure 7).

provided incomprehensible or nonsensical answers (e.g., "Element is Fine text"). We rejected these submissions and obtained new feedback. From the pool of 540 feedback items, we removed four peripheral feedback items that did not target the design itself, e.g., "Why is the open boathouse restricted to only people with a university Harvard affiliations?".

After filtering out invalid feedback, the first two authors of this paper grouped the feedback items that targeted the very same aspect of the flyer design and arrived at the same conclusion. For instance, as shown in Figure 5 (bottom), the three statements target the same visual element, but only the conclusion of the first and second are the same. Therefore, we grouped the first two but not the third feedback item. For each group, we randomly selected only one item. We used these groupings to avoid presenting the same critique multiple times. While the number of identical feedback items can provide an estimate for the critique's severity, we opted for diverse feedback instead. Finally, we randomly selected five feedback items per design from the selection of unique feedback items, which were then shown to the participant during the second session. Given the time constraints for the revision task, we chose to limit the number of feedback items so that the participants did not have to spend much time on organizing the feedback.

5.6 Measurements

We used the results of three survey questions related to the feedback's thought-provokingness, usefulness, and tone as measures for the acceptance of feedback (**H-Support 3**). See Supplementary Figure S17 for an example.

Thought-provokingness. In the second session, after having read each feedback item, but before submitting the revised design, we asked the participants: "Does this [statement/question] make you think about your design in a new way?". The participants provided their answers on a 5-point Likert scale ranging from "no, not at all" (1) to "yes, very much" (5).

Usefulness. After the participants submitted their revised designs, we showed them the feedback again with the original and revised flyer design. This time, the participants had to rate each feedback item's usefulness in regards to the design revision by answering "Was this feedback useful for revising your design?" using a 5-point Likert scale ranging from "no, not at all" to "yes, very much". Our goal was to find out which feedback was perceived useful for revising the design as an indicator of the feedback acceptance.

Tone. We also asked the participants to rate the tone of the feedback on a 5-point Likert scale from "very negative" to "very positive" to get a subjective rating of the feedback's sentiment polarity. To indicate that the tone is different from the feeling, we additionally asked the participants how the feedback made them feel.

Improvement. To assess the impact of the feedback on the design revision (**H-Main**), we asked the jury members to rate the improvement of each flyer design on a diverging 7-point Likert scale ranging from "worsened significantly" (1) to "significant improvement" (7).

5.7 Results

The 36 participants created a total of 72 flyer designs (two designs per participant). Figure 7 shows a diverse sample of eight flyer designs created by the participants. The distributions of key measures per condition (**S**, **Q**, and **Q+S**) are shown in Figure 6.

To assess the overall effect of the feedback conditions on the quality of the design revisions, we analyzed the experts' improvement ratings of the revised flyers. The distribution is shown in Figure 6 (right side). A Kruskal-Wallis rank sum test with condition (**S**, **Q**, and **Q+S**) as the independent variable and improvement as the dependent variable shows a significant effect of the conditions ($H=7.34$, $df=2$, $p=.0255$). A pairwise post-hoc Dunn test with Benjamini-Hochberg correction was significant for **Q+S** versus **S** ($p=.0341$) and **Q+S** versus **Q** ($p=.0479$). However, **S** does not significantly differ from **Q** ($p=.89$). The results show that the mean improvement for **Q+S** ($M=4.77$, $SD=1.36$) was significantly greater than the mean improvement for **S** ($M=4.41$, $SD=1.14$, $d=.29$) and **Q** ($M=4.32$, $SD=1.26$, $d=.34$). The effect sizes for these analyses ($d=.29$ and $d=.34$) were found to exceed Cohen's [9] convention for a small effect ($d=.2$).

A Kruskal-Wallis rank sum test with the condition (**S**, **Q**, and **Q+S**) as the independent variable and thought-provokingness as the dependent variable shows a significant effect of the condition ($H=10.17$, $df=2$, $p=.0061$). A pairwise post-hoc Dunn test with Benjamini-Hochberg correction was significant for **S** versus **Q** ($p=.0079$) and **Q+S** versus **Q** ($p=.0232$). The results show that the mean thought-provokingness of **S** ($M=3.80$, $SD=1.22$) and **Q+S** ($M=3.57$, $SD=1.25$) were significantly higher than **Q** ($M=3.03$, $SD=1.33$). However, **Q+S** did not significantly differ from **S** ($p=.56$). Apart from that, we found no significant effect of condition on either usefulness ($H=3.62$, $df=2$, $p=.16$) or tone ($H=1.75$, $df=2$, $p=.42$).

To determine whether the feedback differed by some other measure, we conducted a Wilcoxon signed-rank test to compare the

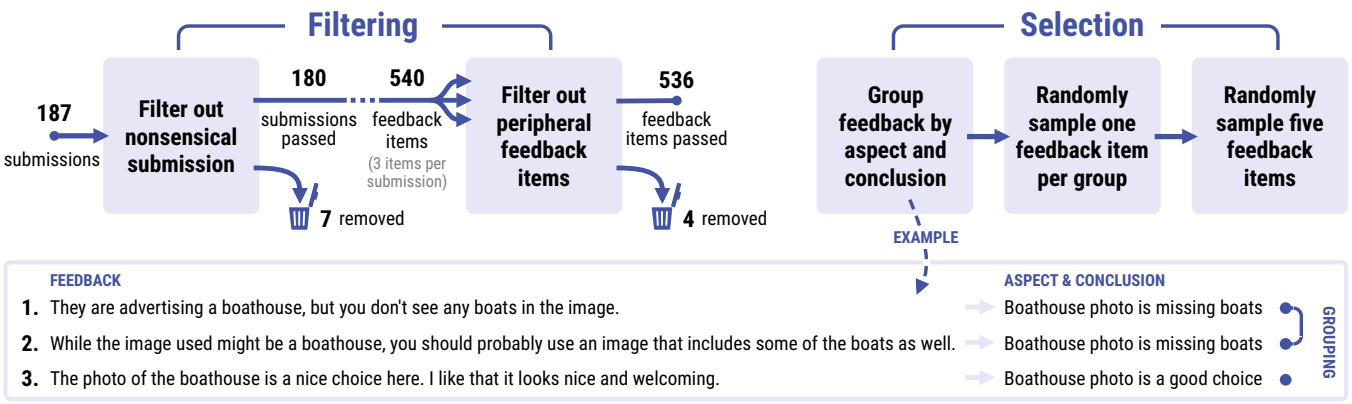

Figure 5: Feedback Selection: First, we rejected nonsensical submissions and removed peripheral feedback items. Next, for each flyer design, we grouped the feedback by the aspect (e.g., font size) and conclusion (e.g., too small) and randomly selected one feedback item per group. From the remaining feedback items we randomly sampled five feedback items that were presented to the participant.

statement length between **S** and **Q+S** and the question length between **Q** between **Q+S**. We found that statements in **S** (M=120.3, SD=52.4) are significantly longer than in **Q+S** (M=87.8, SD=45.0; W=383.0, p<.0001). In contrast, the question length in **Q** (M=90.2, SD=47.5) did not differ significantly **Q+S** (M=90.9, SD=47.0; W=835.5, p=.88). We also compared the feedback's absolute polarity using a Wilcoxon signed-rank test but did not find any significant differences in the statements between **S** (M=.36, SD=.29) and **Q** (M=.33, SD=.35; W=781.5, p=.56) and the questions in **Q** (M=.15, SD=.21) and **Q+S** (M=.2, SD=.27; W=341.5, p=.17).

To verify if the redesigns were based primarily on the feedback obtained through this study, we asked participants after the study: "Did you collect feedback or ideas for the revision elsewhere?" (1 = "no, not at all" to 5 = "yes, very much"). On average, the participants reported that they did not collect ideas elsewhere (M=1.39, SD=.99), and there was no significant difference between the conditions with respect to this question.

6 OVERALL DISCUSSION

Enhancing Feedback With Open-Ended Questions. In terms of the overall effect of **S** (statements only), **Q** (questions only), and **Q+S** (question-based feedback followed by statement-based feedback) on the quality of design revisions, we found that **Q+S**led to significantly better revisions than either **S** or **Q**, which provides evidence in support of our main hypothesis (Table 1). Even though the statement-based feedback we collected lacked strong sentiment on average, the effect sizes of **Q+S** compared to **S** (d=.29) and **Q+S** compared to **Q** (d=.34) show a clear impact on the overall effectiveness of design feedback. Such impact was not evident in previous work on enhancing crowdsourced design feedback [18, 30, 32], which instead focused on improved feedback perception. The improvement in design iteration that we saw might in part be due to the reflective nature of question-based feedback. In this regard, our work extends the findings from Yen et al. [54], who demonstrated that a reflective activity alone can be as effective as feedback for design iterations. Yet, their results did not show a benefit of combining the reflective activity with traditional feedback, which was the case for **Q+S** in our study. Overall, we assume that the impact of **Q+S** will be even greater in contexts where the crowdsourced feedback contains stronger sentiment, such as in social networks or web forums [53].

Furthermore, as expected, we found that feedback in the form of questions only (**Q**) led to the least-improved design revisions. These results, albeit the difference between **S** and **Q** was not significant, are in line with previous work [3] and suggest that question-only feedback should not replace statement-based feedback for novices.

In support of our approach, through manually coding questions as either open-ended and thought-provoking or not, we show that it is indeed possible to enable online crowd workers to rephrase their statements into open-ended and thought-provoking questions. In total, 85% of all questions were successfully rephrased, which we believe is a strong indicator that our AMT task design is an effective approach to crowdsource question-based feedback. Therefore, **H-Support 1** is supported. To further improve the success rate, future work could guide the elicitation of question-based feedback with natural language processing towards open-endedness.

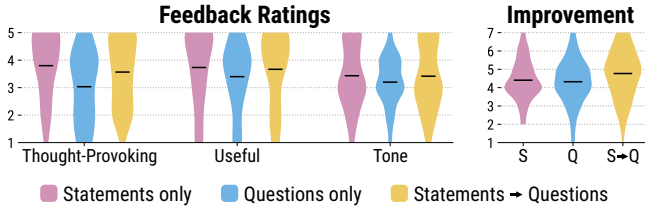

Figure 6: Feedback Ratings and Design Improvements: Distribution of the feedback ratings from the participants and improvement ratings of the jury. Note, the improvement score is provided on a diverging 7-point Likert scale where 1 refers to "worsened significantly" and 7 refers to "significant improvement".

Figure 7: Flyer Designs: Eight flyer designs from study 2. The top row shows flyers with decreasing average quality scores of the revised design. The bottom row shows flyers with decreasing average improvement scores. Each pair of images shows the original design on the left and the revised design on the right. The first flyer (1) won the best design award.

The results of the polarity analysis strongly indicate that questioning is an effective technique to neutralize sentiment. In particular, the sentiment of negative statements is resolved entirely, which is essential to avoid negatively influencing the recipient's affective state. Interestingly, the sentiment of positive statements is also reduced, which suggests that question-based feedback carries less sentiment overall. In conclusion, our results suggest that **H-Support 2** is supported. By presenting question-based feedback prior to statement-based feedback, our method is an implementation of Wu et al.'s approach for mitigating unwanted effects of negative sentiment [48].

Regarding the effects of questions on the perception of statements with overly positive or negative sentiment, we did not find any significant differences between the conditions in the reported usefulness ratings. Therefore, we cannot confirm **H-Support 3**. In comparison, related work [18, 30, 32] found that structuring and scaffolding can improve the feedback's perceived usefulness. A potential explanation why we still saw an improved effectiveness of the **Q+S** feedback compared to **S** and **Q** could be that preposed question-based feedback primarily changes the recipient's focus from themselves to the design task. This change might have mitigated the effects of negative feedback [37]. Contrary to our expectations, the only significantly different feedback rating was thought-provokingness, which was the lowest in **Q**. In hindsight, asking participants about the magnitude of how much a feedback item made them think about their design might have been too unspecific. For instance, instructional feedback could have prompted the participants to think a lot about how to execute suggestions rather than to think about alternative designs. A more in-depth analysis of the revised designs could uncover which feedback was indeed addressed. It might also be necessary to study this question by limiting the feedback to highly negative and positive statements to emphasize the potential effect of questions on the perceived usefulness.

Hypothesis		Support
H-Main	Feedback presented as questions followed by statements improves design revisions compared to statement-based or question-based feedback alone.	✔ Yes
H-Support 1	Non-expert crowd workers can ask open-ended and thought-provoking feedback questions.	✔ Yes
H-Support 2	Question-based feedback has more neutral sentiment than statement-based feedback.	✔ Yes
H-Support 3	Feedback presented as questions followed by statements leads to more balanced acceptance of subsequent statement-based feedback.	✖ No

Table 1: Key Findings: The results support our main hypothesis and two out of three supporting hypotheses.

Generalizability. Given the breadth of related work, we would assume to see similar effects of question-based feedback in other domains. In particular, question-based feedback should easily be applicable to different areas of creative work due to the similar processes of iteration. Regarding our method for crowdsourcing question-based feedback, there are no technical limitations to expanding this method to other types of work. However, the success of crowdsourcing question-based feedback depends on the accessibility of the work to non-expert crowd workers. While graphic design in general and flyer-based advertisement in specific should be accessible by most people, this might not be the case for other types of work.

Beyond crowdsourcing, questions could also be employed as a generic method to enhance feedback. However, the usefulness of question-based feedback might be limited by the ability of the feedback providers to ask effective questions. More work needs to be done to better understand how the effectiveness of questions and statements are related when the feedback is obtained in other contexts, for instance, from domain experts.

Limitations. On average, the design revision improvement across all conditions was in line with previous work on the effectiveness of crowdsourced feedback [30]. However, by splitting the second study into two separate sessions, we might have lowered the participants' motivation and excitement, as they were compensated only after completing both sessions. An effort-based compensation approach might have helped to increase the participants' motivation.

In this study we focused on the feedback's effectiveness for design iteration. In terms of the perceived feedback quality, we did not find any differences except for the thought-provokingness. And while the statement lengths differed between **S** and **Q+S**, it is unclear how to interpret the comparison given that **Q+S** additionally included the questions. One option to generically quantify the quality could be to ask designers to enumerate revision ideas prior to the actual redesign, which we leave as an idea for future work.

More fundamentally, assuming that the statements and questions are of the same quality, questions can reduce the sentiment of feedback statements and potentially facilitate reflection, but they cannot make the feedback, as a whole, more substantive.

7 CONCLUSION AND FUTURE WORK

In this study, we empirically compared the effectiveness of crowdsourced design feedback on design revisions when presented as statements, questions, and a combination of both. Our results show that the combination of question- and statement-based feedback leads to better design revisions. We believe that these findings are generalizable to other kinds of creative work beyond graphic design. Also, we regard presenting feedback as open-ended questions to be complementary to other approaches for improving crowdsourced feedback. Therefore, it can be integrated into existing online feedback systems to improve the overall effectiveness of crowdsourced feedback further.

Future studies may analyze how exactly questions influence the perception of related statements by exclusively examining feedback that carries strongly positive and negative sentiment, or explicitly letting the designer elaborate on their revision to relate changes to specific feedback items. Moreover, it would be interesting to evaluate what aspects determine the quality of question-based feedback regarding reflection. We assume that, similar to statements, the ability of questions to generate productive ideas for design revisions depends on their specificity. However, more aspects likely come into play. Also, given that designers with varying expertise make sense of and provide feedback differently [14, 16], it would be interesting to determine if question-based feedback is perceived differently by non-professional and professional designers.

ACKNOWLEDGMENTS

We would like to express our gratitude to Humphrey Obuobi for his help with the pilot study. Also, we thank all the participants who took part in our user studies. This research was supported in part by a gift from Adobe Research. The second author is partially funded by an Onassis Scholarship (Scholarship ID: F ZO 002/1 – 2018/2019).

REFERENCES

[1] Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology* 5, 4 (2001), 323–370.
[2] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems (DIS '14)*. ACM, New York, NY, USA, 93–102.
[3] Inneke Berghmans, Nathalie Druine, Filip Dochy, and Katrien Struyven. 2012. A facilitative versus directive approach in training clinical skills? Investigating students' clinical performance and perceptions. *Perspectives on medical education* 1, 3 (2012), 104–118.
[4] Caroline Brandt. 2008. Integrating feedback and reflection in teacher preparation. *ELT journal* 62, 1 (2008), 37–46.
[5] Paul Cairns, Pratyush Pandab, and Christopher Power. 2014. The influence of emotion on number entry errors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2293–2296.
[6] Douglas Carnine, Candy Stevens, Jean Clements, and Edward J Kameenui. 1982. Effects of Facultative Questions and Practice on Intermediate Students' Understanding of Character Motives. *Journal of Reading Behavior* 14, 2 (1982), 179–190.
[7] Kristine L Chambers and Joan N Vickers. 2006. Effects of bandwidth feedback and questioning on the performance of competitive swimmers. *The Sport Psychologist* 20, 2 (2006), 184–197.
[8] Angelo V Ciardiello. 1998. Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy* 42, 3 (1998), 210–219.
[9] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (3 ed.). Lawrence Erlbaum Associates, Hillsdale, NJ, USA.
[10] Adam Connor and Aaron Irizarry. 2015. *Discussing Design: Improving Communication and Collaboration Through Critique.* O'Reilly, Sebastopol, CA, USA.
[11] Corel. 2020. Gravit Designer. https://designer.io. Accessed: 2020-02-16.
[12] Patrick A Crain and Brian P Bailey. 2017. 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 (C&C '17)*. ACM, New York, NY, USA, 80–92.
[13] Deanna Dannels, Amy Housley Gaffney, and Kelly Norris Martin. 2008. Beyond Content, Deeper than Delivery: What Critique Feedback Reveals about Communication Expectations in Design Education. *International Journal for the Scholarship of teaching and Learning* 2, 2 (2008), n2.
[14] 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.
[15] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2807–2816.
[16] Eureka Foong, Darren Gergle, and Elizabeth M Gerber. 2017. Novice and Expert Sensemaking of Crowdsourced Design Feedback. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.
[17] Google. 2020. Google Drawings. https://docs.google.com/drawings. Accessed: 2020-02-16.

[18] 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.

[19] Thomas Grill. 2017. Python implementation of Krippendorff's alpha. https://github.com/grrrr/krippendorff-alpha/. Accessed: 2020-02-16.

[20] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research* 77, 1 (2007), 81–112.

[21] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. 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 (CHI '16)*. ACM, New York, NY, USA, 458–469.

[22] Clayton J Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media (ICWSM '14)*. AAAI, Menlo Park, CA, USA.

[23] Hyeonsu B Kang, Gabriel Amoako, Neil Sengupta, and Steven P Dow. 2018. Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–13.

[24] Alison King. 1990. Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal* 27, 4 (1990), 664–687.

[25] Alison King. 1992. Facilitating elaborative learning through guided student-generated questioning. *Educational psychologist* 27, 1 (1992), 111–126.

[26] Alison King. 1995. Inquiring minds really do want to know: Using questioning to teach critical thinking. *Teaching of Psychology* 22, 1 (1995), 13–17.

[27] Cyril H Knoblauch and Lil Brannon. 1984. *Rhetorical Traditions and the Teaching of Writing*. Boynton/Cook Publishers, Upper Montclair, NJ, USA.

[28] 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.

[29] Kurt Luther, Amy Pavel, Wei Wu, Jari-lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2014. 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 (CSCW '14)*. ACM, New York, NY, USA, 21–24.

[30] 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.

[31] 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.

[32] Tricia J Ngoon, C Ailie Fraser, Ariel S Weingarten, Mira Dontcheva, and Scott Klemmer. 2018. Interactive Guidance Techniques for Improving Creative Feedback. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 55.

[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.

[34] Myle Ott. 2020. Unique Turker. https://uniqueturker.myleott.com. Accessed: 2020-02-16.

[35] 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.

[36] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (1989), 119–144.

[37] Joan Sargeant, Karen Mann, Douglas Sinclair, Cees Van der Vleuten, and Job Metsemakers. 2008. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education* 13, 3 (2008), 275–288.

[38] Joan M Sargeant, Karen V Mann, Cees P Van der Vleuten, and Job F Metsemakers. 2009. Reflection: a link between receiving and using assessment feedback. *Advances in Health Sciences Education* 14, 3 (2009), 399–410.

[39] Donald A Schön. 1984. *The Reflective Practitioner: How professionals think in action*. Vol. 5126. Basic Books, New York, NY, USA.

[40] Donald A Schön. 1985. *The design studio: An exploration of its traditions and potentials*. RIBA Publications for RIBA Building Industry Trust, London, UK. 99 pages.

[41] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. 2015. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 937–945.

[42] 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, ACM, New York, NY, USA, 83–92.

[43] Toyin Tofade, Jamie Elsner, and Stuart T Haines. 2013. Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education* 77, 7 (2013), 155.

[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.

[45] Belkis Uluoğlu. 2000. Design knowledge communicated in studio critiques. *Design Studies* 21, 1 (2000), 33–58.

[46] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 227–236.

[47] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '11)*. ACM, New York, NY, USA, 143–146.

[48] Y Wayne Wu and Brian P Bailey. 2017. 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 (C&C '17)*. ACM, New York, NY, USA, 137–147.

[49] Y Wayne Wu and Brian P Bailey. 2018. Soften the Pain, Increase the Gain: Enhancing Users' Resilience to Negative Valence Feedback. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–20.

[50] 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 (CHI '12)*. ACM, New York, NY, USA, 295–304.

[51] 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.

[52] Anbang Xu, Huaming Rao, Steven P Dow, and Brian P Bailey. 2015. A classroom study of using crowd feedback in the iterative design process. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (CSCW '15)*. ACM, New York, NY, USA, 1637–1648.

[53] Yu-Chun Yen, Steven P Dow, Elizabeth Gerber, and Brian P Bailey. 2016. 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 (DIS '16)*. ACM, New York, NY, USA, 773–784.

[54] 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.

[55] 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.