Aligning Crowdworker Perspectives and Feedback Outcomes in Crowd-Feedback System Design

SASKIA HAUG, Karlsruhe Institute of Technology, Germany
IVO BENKE, Karlsruhe Institute of Technology, Germany
ALEXANDER MAEDCHE, Karlsruhe Institute of Technology, Germany

Leveraging crowdsourcing in software development has received growing attention in research and practice. Crowd feedback offers a scalable and flexible way to evaluate software design solutions and the potential of crowd-feedback systems has been demonstrated in different contexts by existing research studies. However, previous research lacks a deep understanding of the effects of individual design features of crowd-feedback systems on feedback quality and quantity. Additionally, existing studies primarily focused on understanding the requirements of feedback requesters but have not fully explored the qualitative perspectives of crowd-based feedback providers. In this paper, we address these research gaps with two research studies. In study 1, we conducted a feature analysis (N=10) and concluded that from a user perspective, a crowd-feedback system should have five core features (scenario, speech-to-text, markers, categories, and star rating). In the second study, we analyzed the effects of the design features on crowdworkers' perceptions and feedback outcomes (N=210). We learned that offering feedback providers scenarios as the context of use is perceived as most important. Regarding the resulting feedback quality, we discovered that more features are not always better as overwhelming feedback providers might decrease feedback quality. Offering feedback providers categories as inspiration can increase the feedback quantity. With our work, we contribute to research on crowd-feedback systems by aligning crowdworker perspectives and feedback outcomes and thereby making the software evaluation not only more scalable but also more human-centered.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: crowdsourcing, feedback, crowd-feedback system, design, experimental study, qualitative interviews

1 INTRODUCTION

The continuous integration of potential users in the evaluation of software is a challenging but critical activity in the design and development process [3]. However, due to their face-to-face character, traditional evaluation methods such as interviews, focus groups, or usability tests lack scalability and are costly. Furthermore, as they are usually conducted with small groups of participants, evaluation results tend to be limited concerning generalizability [23].

Authors' addresses: Saskia Haug, saskia.haug@kit.edu, Karlsruhe Institute of Technology, Kaiserstrasse 89-91, Karlsruhe, Germany, 76133; Ivo Benke, ivo.benke@kit.edu, Karlsruhe Institute of Technology, Kaiserstrasse 89-91, Karlsruhe, Germany, 76133; Alexander Maedche, alexander.maedche@kit.edu, Karlsruhe Institute of Technology, Kaiserstrasse 89-91, Karlsruhe, Germany, 76133.

Leveraging crowdsourcing in software development has received growing attention in research and practice. Commercial platforms like UserTesting, uTest, UserZoom, and UserCrowd offer different forms of crowdsourced evaluation services. In recent years, two research streams have emerged that have the goal to overcome the limitations of traditional software design evaluation forms through crowdsourcing: crowd testing and crowd feedback. Both focus on using the crowd to involve users in software development but differ in their objectives. Crowd testing has the goal to identify system errors and follows existing testing methods like usability testing [17]. Crowd feedback aims to collect individual opinions and perceptions of the software design by users, anonymous crowdworkers, students, or friends and family [12]. It is rooted in the field of visual design where peer feedback is an established approach to iterate design solutions [39]. Since it is not required for crowd feedback to have a high-fidelity prototype, but user stories or screenshots are sufficient, the application of crowd feedback is broader and more flexible. Moreover, crowd feedback is applicable throughout the whole software lifecycle and enables designers to collect diverse feedback in terms of type and scope [12].

Previous research proposed crowd-feedback systems that include various design features and can be applied in a diverse set of contexts. One of the most popular systems is CrowdCrit [20]. CrowdCrit mainly relies on qualitative feedback that users can add to predefined feedback categories. Additionally, users can apply markers to indicate which element or area their feedback is addressing. Thereby, CrowdCrit is mainly designed to evaluate static designs, like posters. There exist only a few crowd-feedback systems that focus on evaluating interactive design prototypes or even software, like AppEcho [34], Critiki [10], and CrowdUI [28].

The majority of existing studies is focusing on demonstrating the feasibility of crowdsourcing feedback in their individual area of application. Thereby, mainly qualitative evaluation has been performed. Only a few studies have investigated the effects of design characteristics of crowd-feedback systems on the feedback quality and quantity in experimental studies following a quantitative evaluation approach. For example, Yuan et al. [47] showed that offering novice crowdworkers feedback categories to indicate on which topics feedback is required has a positive impact on the feedback quality. Other studies showed the effects of very specific characteristics and requirements of the feedback like using a critique style guide [16], framing feedback as questions [18], or viewing the design on which the feedback shall be collected as part of a narrative [43]. However, there exist many different design features of crowd-feedback systems that are frequently applied. These include, but are not limited to questionnaires, free text fields, categories, selection, direct manipulation, recordings, collaboration, markers, and scenarios [12]. However, their individual effects on feedback quality and quantity are not well understood. This represents an important first research gap for the design of crowd-feedback systems.

Additionally, existing studies mainly focused on understanding the requirements of feedback requesters (i.e., designers), but fail to consider the perspective of feedback providers (i.e., crowdworkers). Oppenlaender et al. [26] addressed this issue by comparing feedback providers' and requesters' feature preferences. However, their evaluation did not study the underlying reasons for users' preferences and did not analyze the resulting feedback outcomes. Additionally, not all insights can be transferred to the evaluation of interactive design prototypes or even software. Krause et al. [16] also included crowdworkers in the evaluation of their critique style guide. Still, the crowdworkers' perspective represents only a minor part of the entire evaluation study. We believe that it is important to include the perspective of feedback providers not only in the evaluation but also in the initial design of crowd-feedback systems. This allows us to align crowdworkers' requirements with the feedback outcomes. Understanding the effects of individual design features on the feedback and the feedback provider will help to adapt crowd-feedback systems better to

their context of use. Thus, designers may be supported in selecting the appropriate design features considering their individual situations. This, in turn, will enable feedback requesters to apply crowd-feedback systems and help make the software development process not only more scalable but also even more human-centered. We identify this as a second major research gap in the field of crowd-feedback systems.

In this paper, we address these research gaps with two studies. In the first study, we conducted initial exploratory interviews to better understand the requirements of feedback providers. We explored how feedback providers perceive crowd-feedback system features and understood how these features should be implemented. Based on these insights, we developed *Feeasy*, a crowd-feedback system [13]. Feeasy includes five key features: (1) a description of a usage scenario of the underlying design prototype to offer feedback providers a context, (2) a speech-to-text feature to add feedback comments via voice, (3) a marker feature to specify the elements of the prototype which the feedback addresses, (4) feedback categories to allocate the feedback comment to a specific category, and (5) a star rating for each category to collect additional quantitative feedback. We, subsequently, conducted an experimental study with Feeasy as an experimental artifact that analyzes the effects of crowd-feedback systems with different design features on feedback quality, quantity, and crowdworker perceptions. The feedback quality is measured via the assessment of UI-design skilled crowdworkers who evaluate each feedback comment in five quality categories (helpfulness, specificity, relevance, sentiment, and objectivity). The feedback quantity is measured via the length of feedback comments. In this study, we applied seven treatment conditions, one for each design feature, one basic treatment with no design features, and one full treatment with all five features combined. To further enhance our understanding of the crowdworkers' perspective on crowd-feedback system features, we conducted additional semi-structured interviews. Our results provide evidence that more design features are not beneficial in all use cases, but applying any design features is better than none. Furthermore, we learned that overwhelming feedback providers might reduce feedback quality and quantity and that scenarios are the favorable design feature when considering the crowdworkers' perspective. With our results, we contribute and extend previous research on crowd-based user involvement in the software development process by analyzing and synthesizing the effects of five crowd-feedback design features and thereby aligning crowdworkers' perceptions with feedback outcomes. Thereby, we aim to allow future crowd-feedback systems not only to be more efficient and effective but also to improve the feedback experience for feedback providers (e.g., crowdworkers).

2 CONCEPTUAL FOUNDATIONS & RELATED WORK

2.1 User Evaluation Methods

Prominent methods to evaluate software designs with users are interviews, focus groups, and usability tests [9, 38]. In general, these methods have in common that the involved designers, domain experts, and end-users have to meet virtually or physically to conduct the software usability and user experience (UX) evaluation. Consequently, these methods lack scalability, are time-consuming, and require monetary resources [9, 33]. One solution for these challenges is leveraging crowdsourcing. Specifically, dedicated crowdsourcing platforms are used to evaluate software design solutions [12]. Crowdsourcing increases the scalability of software evaluation and reduces the effort for software developers and designers through its low-barrier accessibility [1, 10]. Additionally, it provides access to a diverse group of people to evaluate the software design [22]. As introduced earlier, the application of crowdsourcing for evaluation purposes can be distinguished between crowd testing and crowd feedback [12]. While crowd testing requests the crowd to conduct tests to identify errors

in a system, crowd feedback asks users for their verbal feedback that includes opinions on and perceptions of a system. Therefore, crowd feedback may be conducted on interactive prototypes, static designs like screenshots and wireframes, and even textual descriptions like user stories. Crowd testing, in turn, requires high-fidelity prototypes that allow for interaction and include the original content of the system. In summary, crowd feedback allows us to intuitively evaluate the entire software design process from user stories to high-fidelity prototypes, and is well suited for this application.

2.2 Crowd-Feedback Systems

There exist multiple systems that support software designers and developers in collecting design feedback on crowdsourcing platforms. Crowd-feedback systems differentiate in form of multiple dimensions [12]. With regards to the subject under investigation, recent crowd-feedback systems focus on collecting feedback on visual designs such as posters [20], specific software applications such as chatbots [4], and websites [28], or mobile apps [34]. Thereby, the systems differ in the phase of the development lifecycle they are focusing on. While some systems focus on collecting feedback during the development process (e.g., [27, 32, 39]), others collect feedback during usage of the software products for further refinements and continuous improvement (e.g., [28, 34, 35]). The collected feedback can mainly be split into two groups: qualitative feedback and quantitative feedback [12]. While qualitative feedback represents mostly texts or videos, quantitative feedback is collected via votes or ratings. The scope of the feedback also differs between existing systems. Most systems collect feedback on non-functional attributes, such as aesthetics and human values. However, the collection of feedback on content and functional attributes is also supported. Similar to other crowdsourcing systems, crowd-feedback systems also differ in the crowdsourcing configuration, which here comprises the type of crowd (anonymous, users, students, convenience) and the incentive (money, involvement and improvement, interest and social compensation, credits, and gamification). Crowd-feedback systems differ also in their design characteristics. Haug and Maedche [12] thereby identified nine design features: questionnaires, free text field, categories, selection, direct manipulation, collaboration, markers, context, and recording. Finally, it has been shown that crowd-feedback systems do not only have positive effects on the process, but also on outcomes such as feedback quality and quantity, and the resulting design.

Crowd-feedback systems provide multiple benefits for software and user interface (UI) designers to continuously evaluate the software designs during the development process. However, they have downsides as well. Design features might provide the ability to collect design feedback focused on dedicated aspects depending on the situation and enable designers to receive high-quality design feedback. Although the feasibility of crowd feedback for various kinds of designs and systems has been proven, there is still a lack of research on the individual effects of specific design features on feedback quality and quantity, as well as the behavior and engagement of feedback providers. Consequently, it remains unclear how and when to apply these features in crowd-feedback systems.

2.3 Crowdworker Perspective in Crowd-Feedback Systems

In summary, one can distinguish two perspectives in crowd-feedback systems: 1) the perspective of feedback requesters that design a system, create crowdworking tasks, and request feedback and 2) the crowdworkers' perspective who conduct the tasks and provide the feedback. Research and practice so far have primarily focused on the development of efficient crowd-feedback systems that generate optimized results for the feedback requester. However, it missed considering the feedback providers' perspective of the crowdworkers, their experience, and their impact on the feedback outcomes.

Robb et al. [30] and Oppenlaender et al. [26] showed that user engagement plays an important role when crowdsourcing feedback. Increasing the engagement of the crowdworkers improves feedback quality and quantity [26, 30]. A potential explanation is the Theory of Interactive Media Effects (TIME) [36]. The TIME states that features, sources, and content of software affect the users' perception as well as their behavior. As a core characteristic, according to the TIME, the interactivity of software features impacts user engagement. The interactivity addresses the methods of interactions that are offered (e.g., clicking, scrolling, dragging). As an explanation, the various interaction methods improve the user's mental representation of the software. As a shortcoming, however, higher interactivity also affords greater perceptual bandwidth and might aggravate efficient usage [36]. The relationship between feature interactivity and the user's absorption in and attitude towards the system is mediated by the ease of use of the software besides its natural- and intuitiveness [36]. Ease of use is an important factor for the success of crowdworking tasks. Therefore, the application of TIME in the context of crowd-feedback systems might allow us to better focus on the crowdworker perspective [36]. While increasing the level of interactivity and subsequently the level of user engagement helps to improve the feedback, better ease of use of the software can be a path towards a higher level of crowdworker experience.

This can also be explained by the concept of information overload [15]. Roetzel [31, p.480] defines information overload as the situation "when decision-makers face a level of information that is greater than their information processing capacity". Being presented with too much information, in our case multiple options to provide feedback, can lead to people failing to respond to inputs or ignoring information [15]. Consequently, when users are overwhelmed by many options, they might ignore some of them or fail to use them. We believe, that there must be a balance between offering multiple modalities of interaction to increase user engagement and presenting too many options and thereby overloading users.

3 STUDY 1: DESIGN OF A CROWD-FEEDBACK SYSTEM BASED ON THE FEEDBACK PROVIDER PERSPECTIVE

The goal of our paper is to design an innovative crowd-feedback system that addresses both, the crowdworkers' and the feedback requesters' perspectives. While increasing the feedback quality and quantity, we aim to provide an enhanced feedback provision experience for crowdworkers. To do so, we conducted a design study, which was already published as a separate poster [13]. In this design study, we derived design principles from literature and evaluated users' experiences with the features in qualitative interviews. Based on the results, we designed and developed the crowd-feedback system *Feeasy*.

3.1 Method

In the design study, we, first, derived an initial crowd-feedback prototype based on existing design features from the literature. Subsequently, we conducted semi-structured qualitative interviews with exemplary design feedback providers (i.e., crowdworkers) after an interaction with a crowd-feedback system (see section 3.1.2). In the following, we present the methodology of this study in more detail.

3.1.1 Procedure. We recruited ten students for the analysis of crowd-feedback system design features. Four participants were female (six male) and they had an average age of 23.10 years (SD = 2.95). We asked for their level of experience with UI and UX design on a five-point Likert scale. Participants reported little experience on average. For the design study, we derived design features from literature and included them in two distinct crowd-feedback artifacts. We split the participants randomly into two groups of five people. The general procedure for both groups was the same. All

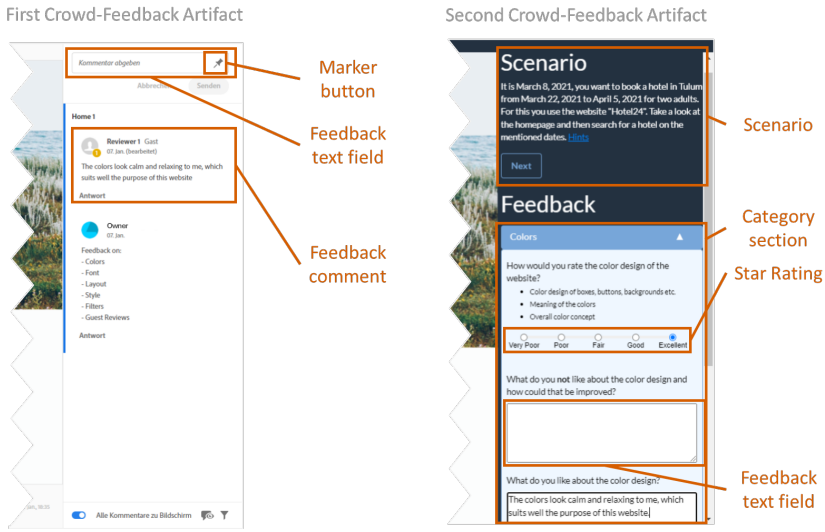

Fig. 1. Screenshots of the two feedback panels for the design study. Left: first crowd-feedback artifact (Adobe XD), right: second crowd-feedback artifact (self-developed).

participants had to interact with one of the two crowd-feedback artifacts to put themselves into the situation of providing feedback and experiencing the design features. Their task was to provide feedback on a low-fidelity prototype of a hotel-booking website. We decided on this prototype because we assume that previous experiences with hotel booking websites among participants are similar. The prototype consisted of four different subpages and blue boxes showed participants where to click. Participants could interact as long as they preferred. Most participants needed 20 - 30 minutes for completing the instructions and the interaction itself. Afterward, they participated in semi-structured qualitative interviews which took around 20 minutes. The qualitative interviews mainly focused on understanding how participants perceived the design features of the crowd-feedback artifacts they interacted with. However, we also asked interviewees about their opinions on further design features that were not included in one of the two crowd-feedback artifacts (e.g., collaboration and voice input). For participation in the whole study, we paid everyone $11.3. The interviews were conducted in German and then translated to English.

3.1.2 Study Artifacts: Crowd-Feedback Systems. We decided to let participants interact with two different crowd-feedback artifacts to be able to receive opinions on multiple design features. Both crowd-feedback artifacts are shown in Figure 1. Design features of crowd-feedback systems can generally be split into nine different types with either the goal to collect feedback (feedback collection mechanisms) or to enrich and improve the feedback (interactivity cues) [12]. These nine design features are free text field, questionnaire, categories, selection, direct manipulation, context, markers, recording, and collaboration [12]. We describe all features in Table 1.

To reduce the development effort in this exploratory phase, we decided to use an existing commercial crowd-feedback system in form of the commenting functionality of the commercialized prototyping software Adobe XD for the first crowd-feedback artifact. This first crowd-feedback artifact collects the feedback in parallel to the design prototype experience. Thereby, the design prototype is on the left side and a panel to add and organize feedback is on the right side. To add a feedback comment, users can enter their feedback in a text field and submit it. As the feedback is

Table 1. Overview of all design features of crowd-feedback systems according to Haug and Maedche [12].

	Feature	Definition	Examples
Feedback Collection Mechanisms	Free Text Field	A single text field for feedback without any specific questions	[34, 43, 46]
	Questionnaire	A series of questions to answer	[1, 25, 27]
	Categories	Categories or rubrics to add feedback comments to	[8, 32, 47]
	Selection	Feedback is provided by selecting items (e.g., rating or voting designs)	[4, 26, 29]
	Direct Manipulation	Design can be edited by feedback providers	[4, 28]
Interactivity Cues	Context	Feedback providers receive a context of use (e.g., a scenario or a persona)	[1, 42]
	Markers	Pins can be put onto the design to indicate which element is meant by the feedback or screenshots/pictures can be added to the comment	[20, 26, 32]
	Collaboration	Feedback providers can interact with the feedback of others (e.g., add comments or vote)	[25, 28, 44]
	Recording	Feedback providers can do voice or video recordings	[7, 26, 34]

only collected via the text field and no categories or questions are included in the UI to guide the users, the free text field is one respective design feature that is applied in this crowd-feedback artifact [12]. After submitting a comment, a new comment box is created. Consequently, all comments are displayed as separate boxes. Thereby, each comment belongs to one subpage of the artifact. Before the study, we added an additional comment that showed users on which aspects feedback shall be provided. Users can also add markers to the prototype to indicate which element their comment is addressing. In general, features that allow feedback providers to annotate the user interface or screenshots by drawing boxes or adding pins (e.g., [20, 35, 42, 45]), help not only feedback provides to feel more engaged [26] but especially support developers in understanding the feedback [34, 35]. The marker feature is according to Haug and Maedche [12] the second design feature of this crowd-feedback artifact.

We complemented the design features in the first crowd-feedback artifact with a self-developed second crowd-feedback artifact that contains further design features. In the following, we outline the design of the second crowd-feedback artifact, which is derived from existing literature, in more detail.

General Layout. The general layout is characterized by the parallel arrangement of the design prototype that allows interaction with a prototype on the left side and the feedback panel on the right. This allows for a close direct connection between the prototype experience and the feedback provision and is innovative compared to other recent crowd-feedback systems in practice (e.g., [27]).

Design Features. The design features which are not covered by the commercial crowd-feedback system that we use in this study, are questionnaire, categories, selection, direct manipulation, recording, collaboration, and context [12]. In the following, we want to provide a short overview of the characteristics of each of these features before explaining which design features are implemented in the second crowd-feedback artifact and why we decided on them. Compared to the free text field, questionnaires ask users specific questions about their perceptions of the design prototype to collect feedback. Usually, each question has a text field, where the crowdworkers can enter

their answer to this question as their feedback (e.g., [44]). Existing crowd-feedback systems apply categories to structure the feedback, guide the feedback providers, and reduce the analysis time of the feedback for requesters as the feedback is already structured [32, 45, 46]. These categories usually represent different dimensions of aesthetics [21], design principles [47], or impressions of the design [45]. Categories can be implemented as narrow statements users can select to add a comment [47] or broader topics that tell feedback providers what kind of feedback is required [32]. The drawbacks of categories are that they might prevent feedback providers from entering feedback that does not fit into these categories [8, 32] or that users might misunderstand the categories and consequently submit wrong feedback. While most studies only use categories as a design element without analyzing their effects, Yuan et al. [47] focused in their study, especially on how categories affect the way people provide design feedback. They learned that categories enable novices to provide feedback that is nearly as valuable as expert feedback. Additionally, they found that this is caused by categories leading to a better writing style. With the selection feature, we summarize all features that enable feedback providers to select something, e.g., a rating score [27], a statement [45] or even a picture [30], to share their feedback. In the educational context, ratings lead to more justifications in the feedback but reduce the feedback quality [14]. Collecting feedback via direct manipulation means that users can adapt the UI or at least some aspects of it according to their wishes to tell feedback requesters how they would like to have it designed. Probably due to the high implementation effort, it is only applied in very few crowd-feedback systems in research (e.g., [28]). The recording feature is usually implemented as video and audio recording of feedback (e.g., [34]). In related studies, it was found that overall, written feedback is more comfortable for feedback providers, but audio recordings could be a helpful alternative [34]. Collaboration in the context of crowd feedback usually means that users can react to the feedback of others by voting or rating it (e.g., [25]). The last design feature, context, includes all features of crowd-feedback systems that provide the crowd with some sort of context in the form of a narrative or a persona that helps them to better understand the context of use. It has shown that offering crowdworkers context increases their empathy and in turn, improves the feedback quality and quantity [24, 39].

We decided to apply categories, selection (in the form of a star rating), and context as design features in the self-developed second crowd-feedback artifact. Combining two feedback collection mechanisms has yet only been done by one other crowd-feedback system [12]. These two design features are easy to combine and do not require a complex implementation such as for direct manipulation. We included in our panel seven category sections, one for each category, to enter feedback. Each section contains two text fields, one for positive and one for negative feedback, and a five-point Likert scale to rate the design aspect.

We decided to apply context as a design feature due to two reasons. In this initial design study, we relied on easy-to-apply and agile development which allowed only simple prototypes. The implementation of context is much easier than developing a recording or collaboration feature. Second, it has shown that offering crowdworkers context increases their empathy and in turn, improves the feedback quality and quantity [24, 39]. However, it has never been analyzed how scenario-based instructions influence feedback compared to simple step-by-step instructions. Therefore, we implemented a scenario that describes users a situation that they should imagine when interacting with the design prototype.

3.2 Results

We analyzed the results of the qualitative interviews deductively by categorizing them. We report them along the categories of general experiences, the design features that were included in the two crowd-feedback artifacts, and further ideas for improvement.

General Experience. The participants appreciated the parallel arrangement of the prototype and the feedback panel to provide comments. Participants in the first group valued the intuitiveness of providing feedback in Adobe XD as it reminded them of the commenting functionality in similar commercial tools (*"It [Adobe PDF reader] is similar with the comments if you make any [they are] also on the right side, so to speak"* (T1P4)). In the second group, participants missed being able to add feedback to one specific subpage (*"I thought that there is basically one feedback for each page and not always one for all"* (T2P3)). Therefore, we derived the implication for crowd-feedback system design of intuitiveness in commenting and specificity for logical subpages. Furthermore, we learned that offering crowdworkers to interact with the design prototype and provide feedback in parallel is highly appreciated.

Scenario. Both groups thought that the guidance through the design prototype by a scenario was helpful to them. In the first crowd-feedback artifact the scenario was not included. However, as participants still needed instructions about where to click, we included the scenario in the overall task instructions for the experiment. Consequently, participants in this group had to jump between the browser tab with the instructions and the browser tab with the crowd-feedback artifact back and forth, which they disliked. In the second group, the participants liked that they could always have an eye on their objective and felt the task was more interactive by having the scenario (*"I found the example at the top very helpful, that you don't just click wildly, because not everything is clickable anyway. And so you had a goal in mind that you can just do, just to test it"* (T2P4)). Consequently, scenarios are a helpful design feature of crowd-feedback systems, as long as they are included in the UI of the crowd-feedback system.

Markers. The markers were perceived as highly positive. Participants in the first group were enthusiastic about the markers as they helped them to be more precise and reduce the risk of being misunderstood by the feedback requester (*"I think I'm a bit more concrete [with my feedback], so there's less room for interpretation"* (T1P5)). In turn, participants in the second group missed an option to directly annotate the prototype and the ability to pinpoint specific elements in connection to their comments. In summary, a marker-like feature was highly requested by participants who did not have the marker feature, while participants of the first group appreciated it as it helped them a lot to focus their feedback.

Categories. While in the second crowd-feedback artifact the categories were included as separate feedback sections, the first artifact showed only a list of the categories. The participants perceived the categories as very helpful in both groups. They reduce uncertainties about the relevance of feedback and point out things that one might have missed otherwise (*"Categories [...] ease it for many people to just start and think about it [their feedback]"* (T2P3)). Some participants mentioned that even more specific categories might be better. The participants who used the first crowd-feedback artifact missed a way to show to which category their feedback comment belongs. Therefore, including categories as sections where crowdworkers can add their comments serves feedback providers as guidance and also helps them to organize their feedback comments accordingly.

Star Rating. Participants in the first group missed *"...something simple, which is quick and from which you can get the necessary feedback"* (T1P3) like a rating or voting functionality. Participants in the second group appreciated the effortless feedback and the ability to combine qualitative and quantitative feedback to offer a broader picture (e.g., *"you first assess that [the design] in itself in these five categories and then you can think more about it"* (T2P1)). Consequently, star ratings seem to be valuable to workers as they offer an effortless way to provide additional feedback besides pure feedback comments.

Further Ideas for Improvement. Since our goal was to evaluate design features in simple and quick prototypes, we did not include all design features for crowd-feedback systems which are relevant based on previous research. Therefore, we asked participants about their opinions on design features that were not included in one of the two crowd-feedback artifacts.

Participants in both groups were indecisive about recording audio comments for their feedback. While some appreciate the reduced time and effort (*"...because it's faster and because I can share my thoughts more quickly instead of having to write them"* (T1P3)), others worried about the reduced structure and mentioned that they feel weird when talking in front of the laptop (*"...when writing, you're more likely to rephrase than when I sit down with a voice recorder and record things"* (T2P4)). Based on this feedback, we also asked them about a speech-to-text feature. While some worried about the accuracy of speech-to-text features, others thought it might be a better solution than a pure recording feature. Consequently, although participants were indecisive about if they would see an overall advantage in using voice input features, especially the speech-to-text feature was appreciated by at least some participants and will therefore be considered in the next iteration of our self-developed crowd-feedback artifact.

Regarding collaboration, most participants agreed that they would get biased when they see what others wrote. They also thought they would feel insecure about sharing unique feedback or think their feedback is useless when others already reported the same ideas. Interviewee T2P5 stated: *"So when you see what others write, you're immediately biased by it. And obviously, it makes it a little bit easier to write your own feedback, but that's not the information that you want to have and that's our job to give you our own feedback."* On the other hand, some participants said that seeing the feedback of others could inspire them to see the design from a different angle. Overall, we think the identified disadvantages of collaboration combined with the higher effort for feedback requesters to handle the collaboration of multiple crowdworkers outweigh their additional inspiration. Consequently, we will not include this feature in further iterations of our crowd-feedback artifact.

From the design study, we know how users perceive selected design characteristics of crowd-feedback systems. Based on these results, we distilled the relevant features and applied the results to design the crowd-feedback system Feeasy.

3.3 Feeasy

Based on the insights of the design study, we iterated our initial self-developed crowd-feedback artifact and developed the crowd-feedback system Feeasy. All features of Feeasy as well as the expected benefits for crowdworkers and feedback requesters are summarized in Table 2. Feeasy is designed to improve both feedback quality and quantity on design prototypes but also aims to improve the crowdworker perspective by increasing interactivity, user engagement, and ease of use for the crowdworkers. In the following, we explain the general layout as well as the individual design features of Feeasy in more detail.

General Layout. Figure 2 shows the final user interface of Feeasy which consists of an interactive design prototype on the left side and a feedback panel on the right side. This layout has been appreciated by the participants of our design study as it allows them to interact with the prototype and provide feedback in parallel. This design feature has the main goal to reduce the effort for crowdworkers which in turn might lead to more feedback. We decided to offer only one text field for users to create new feedback comments. New comments are then added to the panel as separate boxes and belong to the subpage of the prototype on which the crowdworker reported the comment. Each box contains a label that indicates the respective subpage. This shall help crowdworkers to

organize their feedback and in turn, make it better understandable for feedback requesters. In the following, we present the five key design features that we want to evaluate in the following studies.

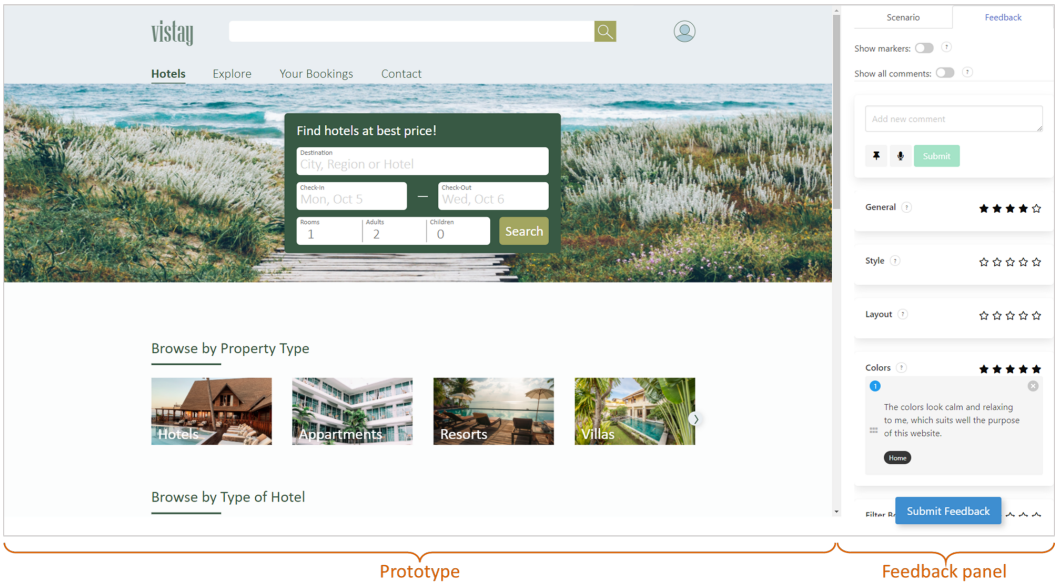

Fig. 2. User interface of the interactive crowd-feedback system Feeasy.

Scenario. In our initial crowd-feedback system, we offered a scenario that told users where to click while providing them with a realistic usage scenario. Participants in the design study saw no disadvantages in having the scenario. Offering feedback providers some sort of context increases their empathy and, in turn, improves the feedback quality and quantity of comments [24, 39]. Therefore, we kept this feature for Feeasy. We decided to move the scenario to a separate tab in the panel to keep the layout simple and clean.

Speech-to-text. Participants in the design study were mainly indecisive about using a voice input feature. In related studies, it was found that overall, text is more comfortable for feedback providers, but audio recording could be a helpful alternative [34]. Consequently, we decided to offer a speech-to-text feature as an optional input mechanism for feedback. The speech-to-text feature enables users to dictate their feedback. When they click on the microphone button Feeasy starts to listen and directly transfers the speech into text. Users can then still edit the text in the text field.

Markers. As markers were found to be helpful for crowdworkers to be more specific and avoid misunderstandings, we implemented them in Feeasy. As already explained, markers help not only feedback providers to feel more engaged [26] but also support developers in understanding the feedback [34, 35]. In our case, the user interface can be annotated with small circles with numbers that belong to one comment box. With the circle, users can indicate which element of the user interface the respective comment is addressing.

Categories. Categories in which users can add respective feedback comments not only enable users to organize their thoughts but also provide value to designers as the collected feedback is already split into categories. It has also shown, that categories help novices to provide better

Table 2. Overview of all features of Feeasy and their potential benefits for feedback providers and requesters.

Feature	Description	Provider Benefits	Requester Benefits
General Layout	Design prototype and feedback panel next to each other	Reduced effort	More feedback
	New separate box for each comment	Providers can organize their comments	Feedback that is already split in separate ideas
Scenario	Textual description of an artificial use case for the prototype	Providers know where to click and get more empathetic	Better and longer feedback
Speech-to-text	Speech-to-text input option in text field	Reduce effort and time for providers while still enabling them to edit and structure their thoughts	Longer feedback with more explanations
Markers	Circles with numbers that can be added to the UI and match the number of one feedback comment	Providers can be more specific with their feedback and avoid misunderstandings	More specific and better understandable feedback
Categories	Sections with headlines in which feedback comments can be added via drag-and-drop	Providers can organize their comments and focus on the aspects on which feedback is required	More relevant and focused feedback and comments already organized in categories
Star Rating	Star rating for each category	Providers have a quick and easy way to share additional feedback	Additional quantifiable feedback and more justifications

feedback [47]. As participants in our design study liked being able to address specific categories and organize their feedback, we kept this feature for Feeasy and just adapted it to the improved layout. We included the categories in Feeasy as separate sections in which comment boxes can be added via drag-and-drop. We decided on categories that mainly focus on aesthetics (layout, color, font, style), and one category that addresses a specific design element (filter bar).

Star Rating. Participants seemed to appreciate the quick and easy way to share feedback with a quantitative evaluation. Additionally, feedback requesters profit from having additional quantifiable feedback that summarizes the qualitative comments. In Feeasy, the quantitative evaluation is included as star ratings. Each star rating is attached to a category. Users can then rate how well they assess each category on a scale from one to five.

4 STUDY 2: EVALUATION OF INDIVIDUAL DESIGN FEATURES OF FEEASY

In our first study, we started by collecting insights on the distinct effects of innovative design features for crowd feedback from the feedback providers' perspective. Based on the results, we designed the crowd-feedback system Feeasy. The primary goal of the second study was to investigate how each individual feature of Feeasy impacts the feedback quality and quantity as well as crowdworker perceptions. Specifically, we compared the individual features with a basic version (no features) and a full version (all features) of Feeasy.

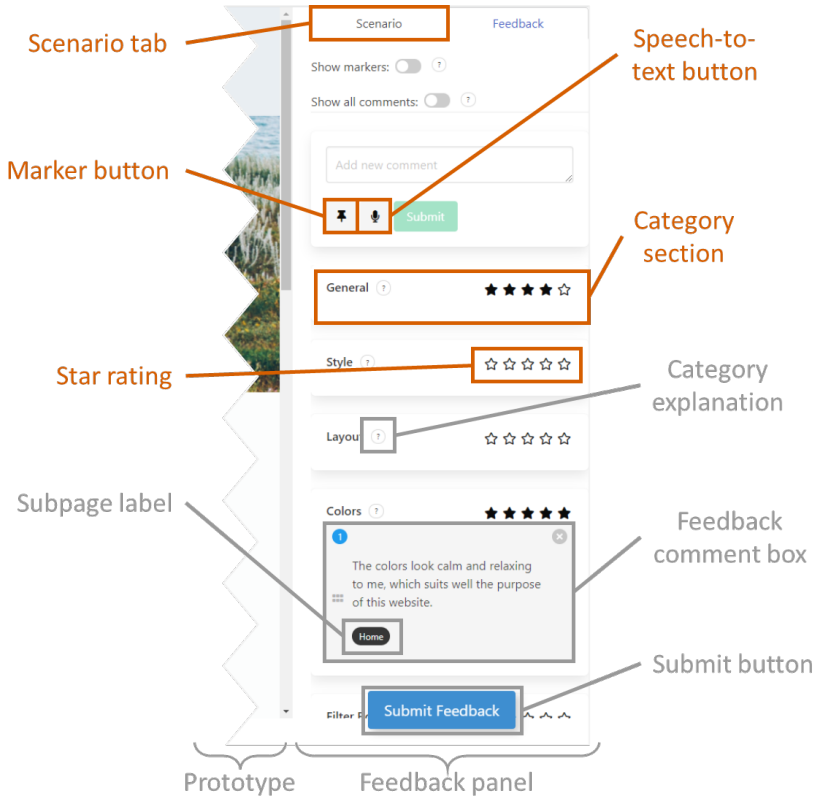

Fig. 3. Feedback panel with an explanation of the general layout (gray) and our five key design features (red).

4.1 Method

To evaluate the individual design features of Feeasy on the crowdworker perceptions in terms of perceived interactivity, user engagement, and ease of use as well as the feedback quality and quantity, we collected design feedback on a fictitious hotel booking website prototype, through a human-intelligence task (HIT) on the crowdworking platform Prolific. Since our goal is to evaluate the effect of each of the five design features individually and, additionally, to compare the results with a baseline and a full version of Feeasy, we derived seven treatment conditions in this study.

4.1.1 Procedure. We implemented seven instantiations of Feeasy: (1) Full (F), (2) Basic (B), (3) Scenario (S), (4) Speech-to-text (R), (5) Markers (M), (6) Categories (C), and (7) Star Ratings (Q). The basic version is displayed in Figure 4. For all five treatments with a single feature, the treatment instantiations looked like the basic version plus the respective feature as implemented in the full version (cf. Figure 3). For example, for the speech-to-text treatment, the Feeasy interface looked like the basic interface (cf. Figure 4) with just the microphone button added below the text field for adding comments. Only for the star rating, we had to additionally include the categories as the rating is always attached to a category. All variants that did not include the scenario contained an interaction tab instead that showed a step-by-step list for each subpage of the prototype to tell users where to click. When starting the HIT, participants received an introduction to the specific version of Feeasy according to the respective treatment as well as a short training on how to provide

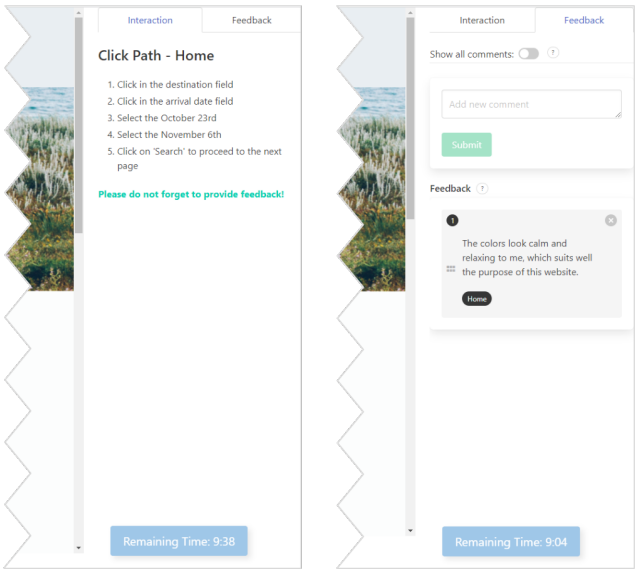

Fig. 4. Basic version of Feeasy without the five key design features.

high-quality design feedback by addressing relevant feedback aspects. After the introduction, each participant was randomly assigned to one of the seven treatment conditions and experienced the treatment phase. During the treatment phase, the participants had to walk through a basic design prototype. Similar to study 1, this was a fictitious hotel booking website (cf. Figure 2). The prototype consisted of four subpages on which the participants gave feedback. To conduct the task the participants had to use Feeasy and provide design feedback comments in the treatment phase for at least ten minutes. After ten minutes they were allowed to submit their design feedback and move on to the next step. The collected design feedback comments, as well as further information (e.g., for which comments the speech-to-text feature was used), were stored in a database. After the treatment phase, participants answered a quantitative questionnaire that asked for their perceived user engagement, perceived interactivity, and perceived ease of use of the experimental prototype Feeasy. Afterward, to additionally collect qualitative data, the participants were offered to book an appointment to participate in an interview. Finally, participants received a debriefing and their compensation.

4.1.2 Participants. For study 2, we recruited 210 participants via Prolific. Of the participants, 48.10% were female (51.43% male) and the average age was 25.70 (SD = 7.62), while the youngest participant was 18 and the oldest 65. On average, the participants reported limited experience in UI/UX design on a seven-point Likert scale (M = 2.30; SD = 1.43). Since the task was to provide design feedback on a hotel booking website prototype, we asked for their frequency of visiting hotel booking websites on a seven-point Likert scale. Their experience with this was limited as well (M = 2.93; SD = 1.24). The participants were distributed on the seven treatment conditions with 29 to 31 participants per treatment. For the task, participants received compensation of $5.0. On top of that, we provided flexible compensation to create a realistic crowdworking scenario and to motivate crowdworkers. The flexible payment was a $1.0 bonus given to participants that ranked within the 30.0% best participants in terms of quality and quantity. Eventually, we paid the bonus to everyone who faithfully completed our task. This resulted in a payment of $6.0 for around

30 minutes of work which is above the German minimum wage ($11.0 per hour). 28 participants took part in the subsequent qualitative interviews, at least three per treatment. The interviews took between 15 and 20 minutes and participants were compensated with an additional payment of $5.0. We removed two of the interviews from the following analysis due to low quality and misunderstandings caused by the language barrier of the crowdworkers.

4.1.3 Data Collection & Analysis. We collected data in two ways. First, we collected quantitative data via the questionnaire for three constructs: perceived interactivity (consisting of fifteen items by [19]), perceived user engagement (consisting of seven items by [40]), and perceived ease of use (consisting out of four items by [5]). For perceived interactivity, we removed the sub-construct synchronicity since all treatments of Feeasy should perform similarly. We also removed all items related to feedback or communication with the website since communication with Feeasy was not relevant to this study. This led us to a final set of seven items for perceived interactivity.

Second, we analyzed the feedback comments collected from the crowdworking task on their feedback quality. Before the analysis, we excluded feedback comments from participants who failed one of our three attention checks (i.e., in form of attention questions in the questionnaire: *"If you are carefully filling out the survey, please select strongly disagree."*). Further, we removed participants that wrote no feedback comments. In the full treatment, we asked participants to rank the five features according to their importance for the feedback. To analyze the feedback comment quality we created another HIT in which UI-design-skilled crowdworkers assessed the quality of the design feedback comments. For this HIT, we again used the crowdworking platform Prolific since it allows us to filter for workers with UI design skills. We recruited 160 workers with UI design experience (M = 4.35, SD = 1.61, based on a 7-point Likert scale). Since the assessment of feedback quality required prior knowledge about the prototype and relevant design feedback dimensions, the participants initially received an overview of Feeasy and the specific aspects they should consider in their assessment.

Subsequently, each feedback comment provided in the initial HIT was analyzed by three participants on the quality categories of helpfulness, specificity, relevance, sentiment, and objectivity. Complementary to the text comment, participants received additional information about potential markers that were added and to which category and subpage the comment belonged. Following previous work on the assessment of feedback quality (e.g., [45, 47]) helpfulness serves as a measure for the overall quality, while the remaining four constructs represent detailed constructs to assess design feedback [16, 27]. A description of each quality construct can be found in Table 3. The feedback quality value for each construct was assessed by taking the average from the distinct ratings of the three individual crowdworkers.

For the qualitative analysis, we conducted semi-structured qualitative interviews with the participants who were willing to provide their insights after the HIT. The questions in the qualitative interviews focused on crowdworkers' experiences with their version of Feeasy in general and each feature in particular. Additionally, we asked participants about their procedure to provide feedback and ideas for further improvement. We analyzed the feedback through a deductive thematic analysis following [2] based on the TIME theory. To facilitate the analysis we organized the results around three categories of general experiences of Feeasy, its positive aspects, and its negative aspects regarding the categories of the TIME theory (i.e., interactivity, engagement, ease of use, feedback quality, feedback quantity).

4.2 Results

4.2.1 Quantitative Analysis. To assess the participants' perceptions, we analyzed the responses to questionnaire items. To assure the internal consistency of latent constructs, we assessed outer

Table 3. Explanation of feedback aspects.

Feedback Aspect	Description
Helpfulness	Helpfulness addresses the overall quality of the feedback comment.
Sentiment	Sentiment assesses if the comment is rather addressing a problem of the design (lower rate) or if it is praising the design (higher rate). Simple statements without judgment should thereby be neutral.
Objectivity	Objectivity evaluates how much the comment is based on facts and not only personal beliefs, opinions, and preferences.
Relevance	Relevance assesses how relevant the comment is to further improve the design of the hotel booking website. Thereby, crowdworkers should consider the categories on which we collected feedback and the limitations of the prototype (e.g., functionalities).
Specificity	Specificity addresses how specifically the feedback has been phrased. This includes how clearly it describes the element it is addressing and its positive or negative aspects.

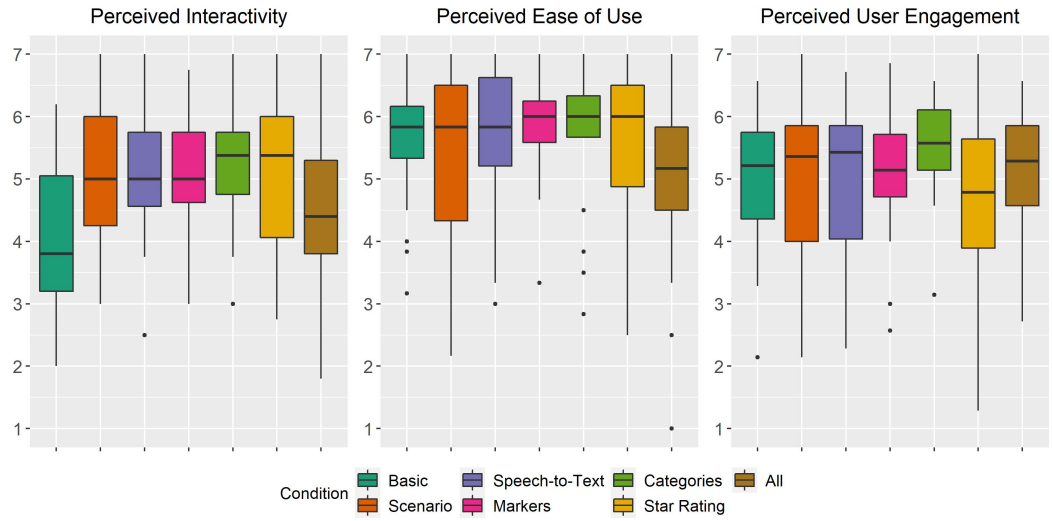

Fig. 5. Boxplots of perceptions of interactivity, user engagement, and ease of use measures of the crowdworkers.

factor loadings and Cronbach's alpha with a cutoff at 0.7 and 0.6 [11, 37]. Since not all constructs did meet these requirements we removed perceived interactivity items five and six having Cronbach's alpha then range from 0.68 to 0.78. Afterward, scales were averaged. To assess the effect of the experimental treatment conditions (basic vs. full treatment), we conducted a multivariate analysis of variance (MANOVA) with the three perceptive measures and the feedback quality and quantity assessments as dependent variables. Since the variables under investigation violated the assumption of univariate and multivariate normality, we conducted a nonparametric rank-based MANOVA using the R software package rankMANOVA (v. 0.0.7) [6]. The results of the rank-based MANOVA for analyzing nonparametric data did not reveal a significant effect of the treatment conditions on the dependent variables. Furthermore, we conducted an aligned rank transform (ART) for nonparametric factorial analyses of variance procedures using the R package ARTool (v. 0.11.1)

[41]. Results show no significant results besides a significant effect of ease of use ($p < 0.05$) between the full and the basic version without design features and a significant main effect for specificity ($p < 0.05$). To complement the quantitative analysis, we, then, pursued a thorough descriptive analysis of the data.

Figure 5 shows that the perceived interactivity for all five treatments with only one feature (except the star rating treatment, which includes two features) is higher than for the basic treatment and the full treatment. Thereby, the perceived interactivity in the scenario and speech-to-text treatment is still lower than for the marker, category, and star rating treatments. The perceived ease of use is in all individual feature treatments similar to the perceived ease of use of the basic treatment and higher than for the full treatment. The results for perceived user engagement differ between the five treatments. While the perceived user engagement for speech-to-text, categories, and the scenario is higher than for the basic and full treatments, the perceived user engagement for markers and star ratings is lower. The highest perceived user engagement was achieved for categories, while the lowest was the star rating treatment, which also included the categories.

The overall quality which was described by the helpfulness of the feedback comment is stable across the individual feature treatments and lower for the basic and especially the full treatment condition. Regarding the sentiment, categories have led to more positive comments compared to the other treatments. However, the differences between the treatments were only marginal. Regarding objectivity, there was no difference between the treatments. All features and combinations of features have led to medium objective feedback comments. The relevance again was the lowest for the basic and full treatment, while there is no difference between the other five treatments. Finally, the specificity is the highest for markers and the lowest for the full treatment. The number of comments per crowdworker was the highest in the category treatment. The lowest number of comments was achieved for the scenario and the full treatment. Regarding the comment length, the results of all treatments were similar with comments having between 70 and 170 characters. Only the number of characters in the full treatment was lower than the rest.

To analyze the results of the ranking task of the full treatment, we calculated Kendall's W to know how much the participants agreed on their ranking. The Kendall-W-Test is a non-parametric statistical test that compares the distributions of three or more related variables and analyzes if these variables are significantly different from one another. Kendall's W can range between 0 (no agreement) and 1 (full agreement). For the test, we transformed the ranking into ordinal values from one to five with one meaning the feature was ranked the most important and five meaning the feature was ranked the least important. We received a Kendall's W of 0.31 which indicated a rather low agreement among the participants. Figure 6 presents stacked bar plots of the rankings of the five design features. The scenario feature was on average ranked the most important ($M = 2.06$) and the recording feature the least important ($M = 4.45$). The star rating and the markers were perceived as similarly important and the categories as slightly less important.

4.2.2 Semi-structured Interviews. In this section, we summarize the insights that we gained during the 26 qualitative interviews that we conducted with crowdworkers who successfully completed the feedback task. We first report their overall experiences and how they proceeded to provide the feedback, then we describe their perceptions of the five individual design features of Feeasy.

Overall Experience. Overall, the participants enjoyed the interaction with Feeasy irrespective of the treatment condition they experienced. All participants in the basic treatment condition appreciated that Feeasy was easy and straightforward to use (*"It's not overloaded with anything, which is great. It's really awesome" (B2)*). One participant even stated that the ease of providing feedback made her "*very willing to give out [...] as much feedback as I could because it wasn't frustrating*" (B3). On the other side, participants in the full treatment condition mentioned the need for more time to

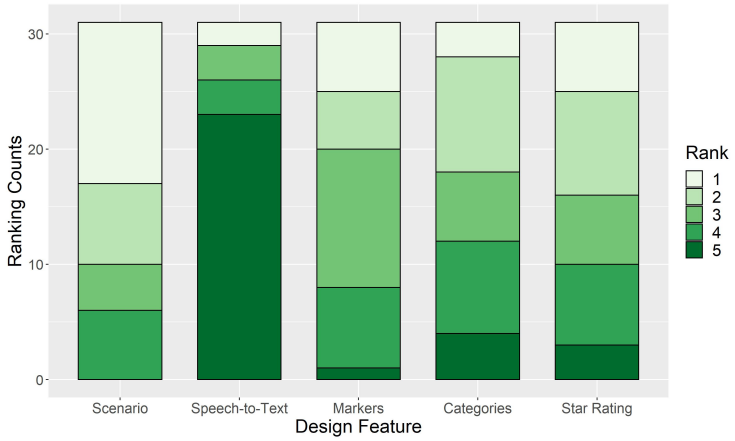

Fig. 6. Stacked barplots of ranking of the five design feedback features (from rank number 1 (best) to rank number 5 (worst)).

get familiar with the system and the options: "*I thought I could maybe have spent a little bit more time looking to give feedback if I've spent less time trying to work out how to work the panel*" (F1). One participant even got frustrated about the interface because s/he didn't understand how to interact with it and stated that s/he would have liked to have a practice before the task to feel more comfortable. In the full treatment condition, participants reported multiple times that they missed using one of the five design features accidentally although they remembered being introduced to the features. For example, F1 mentioned, *"It wasn't immediately apparent in the panel that that [the speech-to-text] was an option. It was just a small icon from what I remember"*.

The remaining aspects that crowdworkers favored or disliked varied a lot between the individual participants and therefore seem not to be related to a specific treatment. Three participants mentioned that they liked the ability to see all previous comments in the feedback panel at once so that they were *"able to keep track of all the comments I've made in the previous time"* (S4). Further participants appreciated, especially in the categories treatment, that they were able to edit their comments after submitting them as they *"...kept finding different stuff that I wanted to add"* (C2). Crowdworkers enjoyed that the system was similar to other systems they use for work and reported that they perceived Feeasy to be interactive and felt engaged by it. A lot of criticism was around the interactivity of the design prototype itself (i.e., the hotel booking website). Crowdworkers stated that being able to click on more things would have led to more feedback (*"If we weren't limited to the testing, if we weren't limited to features, I think that would have improved our feedback results"* (R2)). The second main point of criticism was about the parallel layout. One interviewee recommended placing the feedback panel somewhere else, as *"...it's kind of like narrow and I couldn't see everything clearly"* (Q2). The interviews showed that crowdworkers followed different approaches to identify design issues and report feedback. Very common was that they put themselves in the shoes of another person (e.g., *"...a generic person"* (F3), *"...their grandma"* (B2)) or reported everything that seemed counter-intuitive to them or did not meet their expectations towards the design prototype. In detail, they often looked *"...for things that were different and similar to websites that I know"* (F3) because *"...if we don't have what to compare, I don't think we can choose what is best, what is worse, what can improve"* (C1). Crowdworkers in treatments that did not include the category section still used the categories that were provided in the instructions to make sure that they addressed every category. Participants without the category feature *"...just reported everything that came to their*

mind" (B3), while participants with categories used them to decide which feedback is relevant for the feedback requester.

Some workers reported that they experienced problems with identifying issues with the prototype's UI or *"...find words to explain what is going wrong on the page"* (M2). To sum up, crowdworkers were very positive about their experience with Feeasy. Overall, they appreciated its simple and intuitive UI (*"I feel like that your feedback box is perfect for every user because it's simple and straightforward"* (R2)).

Table 4. Summary of crowdworkers' perspectives on the design features derived from the interviews.

	Advantages	Disadvantages
General Layout	All submitted comments visible Comments can be edited Similar to other tools Interesting, interactive, and engaging	Only one scenario/click path included Lack of guidance
Scenario	Goal-oriented Better focus Equalizes previous knowledge Interaction more real Clear and straightforward	Feedback focused on click path Lower readability Hard to understand
Speech-to-text	Inclusive Higher quality (more comprehensive) Less time and effort More feedback	Option not clear Slower than typing Less organized Feedback more casual Feedback less reliable Not always convenient
Marker	Easier Comments more specific and detailed Small items highlighted	Redundant Feedback too specific
Categories	Better organization Inspiration and guidance Feedback more complete Better focus	"General" too general Moving comments is annoying Less generic comments No methodological guidance
Star Rating	Flexible and easy Good summary of comments Relativizes harsh feedback	Additional effort

In the following, we present detailed insights for each of the five design features.

Scenario. Crowdworkers appreciated that the scenario feature was clear and straightforward and provided them with a goal to focus on. However, the interviews showed that crowdworkers did not perceive the scenario as a design feature. Crowdworkers felt that the scenario made their feedback more real and relevant to the designer (*"With that text, we can give better feedback because we imagine ourselves like these people like we are going to travel"* (S3)). Additionally, they liked that they knew on which parts of the user interface they should focus (*"Maybe it somehow points my attention to specific things. That might have been helpful"* (S5)). One interviewee even stated that the scenario might be especially helpful *"...for not so experienced travelers or new travelers"* (S3). However, F1 reported problems with the scenario instructions since s/he did not find all subpages of the design prototype and finally gave up. Some participants also would have preferred bullet points

instead of a block of text to make it more readable. Participants with the step-by-step instructions did not report any problems, however, they mentioned the creative limitation of the restriction to one user flow (*"I realized that even if I was tempted to play around a bit with the website, I needed to focus on the goals. So, my focus was on actually completing the steps even though [...] you just automatically want to just hover over the little things and see what is what."* (B1)).

Speech-to-text. Our log data showed that none of the participants actively used the speech-to-text feature. Nevertheless, due to its sole presence, most interviewees saw its advantages as *"...your spoken word is better than written text"* (F1). Therefore, we asked crowdworkers about their reasons for not using this feature and what advantages they still see in entering feedback via speech-to-text. Participants provided various reasons why they did not use it: they were not drawn to it (F1), they did not use it because they did not want to disturb the people around them (F3), or they expected the speech-to-text feature to malfunction because of their accent or the quality of their voice and feared to have to recheck all the feedback as the speech-to-text feature might misunderstand them (*"I'm sure it wouldn't catch anything"* (R2)). Furthermore, crowdworkers expected their feedback to be less organized and more casual when using the speech-to-text feature (*"I feel like maybe when I type, I'm more formal in my phrasing than if I was speaking"* (R1)). Interviewee R3 did not use the speech-to-text feature because s/he assumed that s/he types much faster than s/he speaks. Interestingly, the other two interviewees of the speech-to-text treatment reported the major advantage of the speech-to-text feature in the reduced time and effort for providing feedback as it corrects the spelling and feels for them to be easier than typing (*"...overall, I can say it's much easier to use than typing"* (R2)). Additionally, it makes the feedback provision process more inclusive as also crowdworkers that have problems with fast typing, for example, caused by a disability, could easily provide feedback. Furthermore, the speech-to-text feature could lead to time savings, as it corrects the spelling and feels for some crowdworkers to be faster than typing. Interviewees also stated that they could imagine that *"...there will be more explanation when I say it vocally than typing"* (R2) and that they *"...probably would have given more feedback"* (F1).

Marker. The prevailing perception of the markers was positive in the crowdworkers' interviews. Participants mentioned occasionally that the use of markers is in some cases redundant (*"If I'm describing icons or the selection menu I think it doesn't require pinpointing with a marker"* (M1)), and that they might sound too focused when using them (*"I didn't want to sound like to focus one particular thing."* (M1)). On the other hand, the comments of the participants got more specific and detailed and the markers made it possible to highlight very small items like icons: *"It allows you to pinpoint the specific areas which creates greater visibility and you know there are more layers on your feedback"* (M3). Interestingly, participants without the markers stated multiple times that they would like to *"...just click on something and it being a reference to my feedback"* (B3) and thereby indirectly mentioned the benefit of the design feature of markers.

Categories. The crowdworkers perceived the categories as beneficial for the feedback provision. They liked that the categories helped them to better organize their feedback (*"I think it was just more concrete and more structured than it would be without it"* (C3)) and used it for inspiration and guidance (*"I don't have to wonder what should more I write. [...] I have something to each topic [...] and then I just kept adding if I found something"* (C3)). They felt like the categories helped them to provide more complete and specific feedback and focus on the important aspects of the evaluation process (*"Knowing it [the filter box] was a focus and the main topic of a category, I was able to spend more time on that and it definitely helped for sure"* (C2)). F3 would have liked to have even more categories to provide feedback to. On the other hand, participants stated that *"...it was quite hard to move comments into the specific subsections"* (F3) which annoyed them. One crowdworker also

had problems with categories as s/he was not sure what kind of feedback was expected from them (*"I don't think they helped me very much in how to analyze"* (C1)). Finally, using categories too extensively might lead to less feedback that addresses general aspects like the overall style of the website (*"If you put too many categories then you risk of focusing too much on these specific things and not focus on the general website and not give complete feedback and comments on how the website looks as a whole"* (C2)). At the same time, interviewees did not like the 'General' category as *"...the 'general' category gets very general. And what does that mean? It's not really very specific"* (F1)).

Star Rating. In the star rating treatment, crowdworkers could use the categories for their feedback and add an additional star rating for each of the categories. The only disadvantage that crowdworkers reported about the star rating was that it was an additional effort compared to providing just a text comment. However, they still stated that they *"...didn't really find it necessary [...] but it was nice because I could sum up what I thought about it"* (Q2). F1 tried to make sure that *"the star rating was compatible and mirrored the feedback that I had given"*. Furthermore, crowdworkers liked that the star rating provided them the flexibility to rate each category differently. They perceived the star rating to be very easy and good for providing a summary of the text feedback (*"I think it's a very quick way to just say 'OK, this is what my general thoughts were'"* (F3)) and relativizing feedback that might sound too blunt as *"...the stars are very international"* (F2). Further, they thought a good star rating might soften very critical feedback, so that the requester understands that despite the criticism the feature is good (*"If I will not do ratings, then nobody will be able to understand how much I really like it and how much I did not like it"* (Q1)).

5 DISCUSSION

In this section, we synthesize and discuss the insights that we gained in our two studies. Thereby, we put the feedback provider (e.g., crowdworker)) perspective in the focus for crowd-feedback system design and highlight the interplay with the feedback requester objectives. Our results provide evidence that more design features are not significantly better than applying no design features at all. Furthermore, we learned that especially scenarios were appreciated by crowdworkers, and single-feature treatments performed better than the full and basic treatments in terms of crowdworkers perceptions, feedback quality, and feedback quantity. Based on these insights we derived design implications for the design of crowd-feedback systems that align both the feedback requester objectives and the crowdworker experience.

5.1 Number of Design Features

According to the TIME theory by Sundar et al. [36] the perceived interactivity of Feeasy should increase when combining multiple features as this offers feedback providers more interaction opportunities. However, in our case, the perceived interactivity increased when applying one feature compared to no features, but it decreased for the full treatment with five features. It also showed that the perceived ease of use is stable across the basic treatment and the treatments for the individual features, but lower for the full treatment. However, as the perceived user engagement is similar across all treatments this could support the statement of Sundar et al. [36], that additional factors such as naturalness, intuitiveness, and ease of use are important mediators for the relationship between perceived interactivity and perceived user engagement.

We also assume the perceived interactivity to have an impact on the resulting feedback quality as the feedback from the basic and full treatments which achieved the lowest ratings for perceived interactivity also performed the worst in terms of feedback quality. This holds in particular with the helpfulness category. We found evidence in our interviews that crowdworkers were overwhelmed by having so many options which can be explained by the concept of information overload [15, 31].

Crowdworkers, therefore, needed some time to get familiar with them. This might have harmed the feedback that these crowdworkers provided. However, we assume the relationship between the three perception constructs and the resulting feedback quality and quantity to be more complex than we expected initially. Based on the insights from the interviews, we believe that additional aspects such as learnability and understanding need to be considered when designing crowd-feedback systems. In the crowdworking context, the simplicity and clarity of tasks and instructions are key.

While the feedback quality for the full treatment is lower than the feedback quality of the single-feature treatments, the full treatment still provides multiple additional benefits to feedback requesters. Most of the feedback is already categorized, some comments include markers that might increase the comprehensibility and in addition, a quantitative assessment is provided. The feedback quantity is slightly lower for the full treatment than for the other variants. This could be caused by the additional effort and time workers had to spend on learning multiple features. As workers spend more time learning the features they had less time to spend on writing feedback comments. Consequently, applying multiple features might have a negative impact on crowdworkers' perceptions, as well as the feedback quality and quantity.

The results of our studies show that a well-dosed application of certain design features has beneficial effects on crowdworkers and their feedback. As the combination of features might decrease the perceived ease of use and therefore negatively impact overall crowdworkers' perceptions, our recommendation is to use only the design features that are necessary to fulfill the requirements of the design evaluation.

5.2 Individual Features

The main goal of study 2 was to compare the effects of the five design features (scenario, speech-to-text, marker, categories, star rating) of our crowd-feedback system Feeasy on crowdworkers' perceptions as well as feedback quality and quantity. Overall, the perceived user engagement of crowdworkers and the helpfulness of the resulting feedback comments did not seem to be directly related to each other. For example, the perceived user engagement in the basic treatment was similar to the perceived user engagement in the marker treatment. However, the helpfulness of the feedback comments in the marker treatment is rated much higher than that of the feedback comments in the basic treatment. The same applies to the relationship between perceived user engagement and feedback quantity. While the perceived user engagement is the lowest for the star rating, the feedback quantity is similar to the other treatments.

Considering this, the independence between feedback quality and quantity and the crowdworkers' engagement in our study may imply that in the crowd-feedback context other factors play a crucial role to increase the feedback quantity and quality. We learned in the interviews, that crowdworkers appreciate clear and easy features as well as structured guidance in performing their tasks. Additionally, we understood that some workers were insecure about the requesters' expectations of their feedback which might have negatively influenced their feedback quality and quantity. Consequently, additional influencing factors on the feedback quality and quantity might be how well users understand their task and how well the system supports them in expressing themselves and guides them through the feedback task. From the feedback requester's perspective, the objective is to receive feedback with high quality in large quantities. Looking at our quantitative results for the feedback quality and quantity, we were not able to identify significant differences between the treatments. Therefore, we will connect the descriptive results with the interview insights in order to understand the effects of the individual features. Participants of the full treatment ranked the scenario feature as the most important for providing feedback. We assume the reason was that they did not understand how to interact with the prototype without knowing where to click. Consequently, they perceived the scenario as essential to provide feedback. It also might

have helped to make the feedback situation seem more natural. When leveraging the scenario, feedback requesters must consider that the scenario leads to workers' feedback being more focused on specific elements and features. Consequently, the scenario is helpful in particular, when feedback is required for a specific part of the design prototype like a new feature. Looking at the crowdworkers' perceptions, the categories feature performed the best for all three perception constructs (perceived interactivity, perceived ease of use, and perceived user engagement). Remarkably, the categories were the only feature in which crowdworkers' reported usability issues in the interviews. Crowdworkers reported that they had problems moving the comment boxes into the right category sections. This means that the lower perceived ease of use did not influence the positive perception of user engagement and interactivity. When looking at the interview results, one of the main issues of crowd-feedback tasks was that crowdworkers were insecure about the focus of the study, the right specificity of their feedback, and had problems with keeping an overview of their feedback. As the category feature addressed all of these problems, crowdworkers felt more secure and used the categories as guidance for the task, which might have covered up the usability issues and in turn led to the high value for perceived ease of use. Regarding the importance of features for feedback, crowdworkers still ranked star ratings and markers higher. While ratings and markers enable feedback providers to enrich their textual feedback with additional feedback, the categories only offer a better structure. For markers, the crowdworkers' experience with the feature matches the quantitative outcomes as the feedback got more specific in this treatment. Comparing the star rating treatment that contained also categories with the treatment with only categories, the user engagement was lower while the feedback quality was higher. The user engagement was even the lowest for the star rating treatment. The reason for this might be the increased complexity of two features that lead to a higher mental workload for feedback providers. Feedback providers ranked the speech-to-text feature as the least important which is consistent with them not using it at all. The interviews and the feature ranking confirmed that they perceive the feature as a nice add-on, but not essential for providing good feedback. Still, the pure presence of the feature had a positive impact on crowdworkers' perceptions and the resulting feedback compared to not having any feature included. In the interviews, workers were not completely averse to using the feature. We assume after workers get familiar with feedback-providing tasks, some would start using the feature. Still, the value and effect of the speech-to-text feature should be analyzed in future studies.

5.3 Design Implications

Based on our results we here provide a summary of design implications for crowd-feedback systems.

Focus on Crowdworkers' Perceptions. Crowdworkers perceived the scenario feature as the most important. Therefore, we recommend providing a scenario in feedback tasks to guide feedback providers. Although the category treatment performed the best in terms of user engagement, markers and star ratings were perceived as more important by crowdworkers. This is consistent with the qualitative results as many workers who did not have a marker feature in their version of Feeasy, asked for a feature to annotate the user interface of the design prototype. For the star rating, this does not apply. Consequently, markers seem to have a bigger positive impact on crowdworkers perceptions of the crowd-feedback system and should therefore be applied additionally to the scenario when aiming to positively impact crowd perceptions.

Focus on Feedback Quality. The feedback quality was the lowest for the full and the basic treatment. Consequently, we recommend applying selected features when designing crowd-feedback systems and paying attention to balancing the advantages of multiple features and the increased complexity. For the single-feature treatments, there is no feature that clearly performed better than the others.

Each feature has individual advantages and feedback requesters must understand their feedback requirements to select the appropriate features.

Focus on Feedback Quantity. When aiming for many feedback comments, feedback requesters should apply categories or markers. Adding star ratings to the categories has only a minor negative impact on the feedback quantity and could therefore also be an option. Regarding the length of feedback comments, scenarios are the best choice, followed by categories (with and without star ratings), and markers. Applying all five design features has a negative impact on the length of feedback. As each feature takes some time to get familiar with it, generally fewer features are beneficial when aiming for many long feedback comments. Regarding the number of comments, categories are the favorable design feature.

6 LIMITATIONS AND FUTURE WORK

While we followed a rigorous evaluation approach several limitations apply to our study. In the following, we provide an overview of limitations and present future research directions.

Relevance of Design Prototypes. First, our crowd-feedback artifact Feeasy was designed for the collection of feedback for all sorts of design prototypes over all phases of the design process. However, in our evaluation studies, we used always the same design prototype to guarantee for comparability of the results. This was necessary since we focused on the evaluation of the design features. Future work should expand the design feature evaluation with additional design prototypes from different design phases. As we learned that workers use their personal expectations and experiences with similar websites to come up with valuable feedback, the workers' requirements for a crowd-feedback system could be much different when the feedback is collected on a less common type of software.

Investigation of the Speech-to-Text Feature. In all of our three studies, no worker has used the speech-to-text feature. Consequently, the reported perceptions and effects on the feedback are only based on workers' assumptions about their interaction with the feature. Additionally, the changes in the feedback quality and quantity are only caused by the presence of the feature. On the one hand, this shows that the sole presence of features has an effect on crowdworkers. On the other hand, we are not able to make statements about how the usage of a speech-to-text feature affects feedback quality and quantity. In our studies, workers reported multiple advantages and disadvantages of a speech-to-text feature. Especially a recording feature would enable designers to consider more factors than just the pure content of the feedback (e.g., tone). Therefore, we suggest future research to study voice input features for feedback individually. The results might also be relevant for other domains like app store reviews.

Interdependencies between Design Features. In this paper, we presented two studies focusing on a specific set of individual design features relevant to crowd-feedback systems. We assume that there exist interdependencies between the individual design features. Especially in the first design study, the perceptions of the participants might be influenced by interdependencies of the individual features. We attempted to counteract this by asking participants specifically about their perceptions of each individual feature.

Still, the analysis of potential interdependencies is beyond the scope of our paper. However, understanding how the combination of design features affects crowdworkers' perceptions and the resulting feedback, might be very valuable for the design of crowd-feedback systems in research and practice. Therefore, future research should expand on an analysis of the interaction effects of design features. This knowledge can be used by feedback requesters (i.e., designers) to design crowd-feedback systems according to the requirements of their feedback studies. To enable feedback

requesters to instantiate these individualized crowd-feedback systems without large effort, a crowd-feedback system configurator would be beneficial. This configurator could guide feedback requesters in creating dedicated crowd-feedback systems that are adapted to their needs. This would enable designers and developers to easily integrate crowd-feedback systems in all phases of their software lifecycle.

7 CONCLUSION

Design features of crowd-feedback systems have an impact on the resulting feedback. While most existing studies in this context focused on analyzing the feedback outcomes for requesters, we aimed to align crowdworkers' perceptions on a spectrum of different design features with quantifiable effects on feedback quality and quantity. We conducted two studies, in which we first developed the crowd-feedback system Feeasy and, subsequently, used it to analyze distinct five design features for crowd-feedback systems. Our results provide evidence that more design features are not beneficial in all use cases, but applying any design features is better than none. Furthermore, we learned that scenarios and markers are favorable design features when considering the crowdworker perspective, while for the feedback quality and quantity, it is primarily important to not overwhelm crowdworkers with too many complex features. Still, the application of any feature improves feedback quality and quantity. We enrich these findings with profound details on the advantages and disadvantages of each design feature as perceived by crowdworkers. Our findings motivate further investigations for the future design and configuration of design features which are combined to achieve specific effects and serve as a basis for the development of a crowd-feedback system configurator. Overall, we contribute with our work to make the software development process not only more scalable but also more human-centered.

REFERENCES

[1] Oshrat Ayalon and Eran Toch. 2019. A/P(rivacy) Testing: Assessing Applications for Social and Institutional Privacy. In Extended Abstracts of the 2019 CHI Conference. Association for Computing Machinery (ACM), New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312972

[2] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[3] Manuel Brhel, Hendrik Meth, Alexander Maedche, and Karl Werder. 2015. Exploring Principles of User-Centered Agile Software Development: A Literature Review. Information and Software Technology 61 (2015), 163–181. https://www.sciencedirect.com/science/article/pii/S0950584915000129

[4] Yoonseo Choi, Toni-Jan Jan Keith Palma Monserrat, Jeongeon Park, Hyungyu Shin, Nyoungwoo Lee, and Juho Kim. 2021. ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW 4, CSCW3 (2021), 19–23. https://doi.org/10.1145/3432924

[5] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly: Management Information Systems 13, 3 (1989), 319–339. https://doi.org/10.2307/249008

[6] Dennis Dobler, Sarah Friedrich, and Markus Pauly. 2020. Nonparametric MANOVA in Meaningful Effects. Annals of the Institute of Statistical Mathematics 72, 4 (8 2020), 997–1022. https://doi.org/10.1007/s10463-019-00717-3

[7] Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In Proceedings of the 2013 Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 227–236. https://doi.org/10.1145/2470654.2470686

[8] Matthew W Easterday, Daniel Rees Lewis, and Elizabeth M Gerber. 2017. Designing Crowdcritique Systems for Formative Feedback. International Journal of Artificial Intelligence in Education 27, 3 (2017), 623–663. https://doi.org/10.1007/s40593-016-0125-9

[9] Anita Gibbs. 1997. Focus Groups. Social research update 19, 8 (1997), 1–8.

[10] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In C and C 2015 - Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition. Association for Computing Machinery (ACM), New York, NY, USA, 235–244. https://doi.org/10.1145/2757226.2757249

[11] Joe F. Hair, Marko Sarstedt, Lucas Hopkins, and Volker G. Kuppelwieser. 2014. Partial Least Squares Structural Equation Modeling (PLS-SEM): An Emerging Tool in Business Research. , 106–121 pages. https://doi.org/10.1108/EBR-10-2013-0128

[12] Saskia Haug and Alexander Maedche. 2021. Crowd-Feedback in Information Systems Development: A State-of-the-Art Review. In Proceedings of the 42nd International Conference on Information Systems (ICIS) 2021. Association for Information Systems (AIS), New York, NY, USA, 17. https://aisel.aisnet.org/icis2021/is_design/is_design/4

[13] Saskia Haug and Alexander Maedche. 2021. Feeasy: An Interactive Crowd-Feedback System. In Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology, UIST 2021. Association for Computing Machinery (ACM), New York, NY, USA, 41–43. https://doi.org/10.1145/3474349.3480224

[14] 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. Association for Computing Machinery (ACM), New York, NY, USA, 458–469. https://doi.org/10.1145/2858036.2858195

[15] Starr R. Hiltz and Murray Turoff. 1985. Structuring Computer-Mediated Communication Systems to Avoid Information Overload. Commun. ACM 28, 7 (7 1985), 680–689. https://doi.org/10.1145/3894.3895

[16] 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. Association for Computing Machinery (ACM), New York, NY, USA, 4627–4639. https://doi.org/10.1145/3025453.3025883

[17] Niklas Leicht. 2018. Given Enough Eyeballs, all Bugs are Shallow - A Literature Review for the Use of Crowdsourcing in Software Testing. In Proceedings of the 51st Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 4102–4111. https://doi.org/10.24251/HICSS.2018.515

[18] 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 Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 12. https://doi.org/10.1145/3411764.3445507

[19] Yuping Liu. 2003. Developing a scale to measure the interactivity of websites. Journal of Advertising Research 43, 2 (2003), 207–216. https://doi.org/10.1017/S0021849903030204

[20] Kurt Luther, Amy Pavel, Wei Wu, Jari Lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven Dow. 2014. CrowdCrit: Crowdsourcing and Aggregating Visual Design Critique. In CSCW Companion '14: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. Association for Computing Machinery (ACM), New York, NY, USA, 21–24. https://doi.org/10.1145/2556420.2556788

[21] 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. Association for Computing Machinery (ACM), New York, NY, USA, 473–485. https://doi.org/10.1145/2675133.2675283

[22] Xiaojuan Ma, Yu Li, Jodi Forlizzi, and Steven Dow. 2015. Exiting the Design Studio: Leveraging Online Participants for Early-Stage Design Feedback. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. Association for Computing Machinery (ACM), New York, NY, USA, 676–685. https://doi.org/10.1145/2675133.2675174

[23] Wendy E Mackay. 2004. The Interactive Thread: Exploring Methods for Multi-Disciplinary Design. In Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '04). Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/1013115.1013131

[24] D Muñante, A Siena, F M Kifetew, A Susi, M Stade, and N Seyff. 2017. Gathering Requirements for Software Configuration from the Crowd. In 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW). IEEE, 176–181. https://doi.org/10.1109/REW.2017.74

[25] Michael Nebeling, Maximilian Speicher, and Moira C Norrie. 2013. CrowdStudy: General Toolkit for Crowdsourced Evaluation of Web Interfaces. In EICS '13 : proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems. Association for Computing Machinery (ACM), New York, NY; USA, 255. https://doi.org/10.1145/2494603.2480303

[26] Jonas Oppenlaender and Simo Hosio. 2019. Towards Eliciting Feedback for Artworks on Public Displays. In C&C '19. Association for Computing Machinery (ACM), New York, 562–569. https://doi.org/10.1145/3325480.3326583

[27] Jonas Oppenlaender, Elina Kuosmanen, Andrés Lucero, and Simo Hosio. 2021. Hardhats and Bungaloos: Comparing Crowdsourced Design Feedback with Peer Design Feedback in the Classroom. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445380

[28] Jonas Oppenlaender, Thanassis Tiropanis, and Simo Hosio. 2020. CrowdUI: Supporting Web Design with the Crowd. Proceedings of the ACM on Human-Computer Interaction 4, EICS (2020), 1–28. https://doi.org/10.1145/3394978

[29] David A Robb, Stefano Padilla, Britta Kalkreuter, and Mike J Chantler. 2015. Crowd Sourced Feedback with Imagery Rather than Text: Would Designers Use It?. In Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems, Vol. 2015-April. Association for Computing Machinery (ACM), New York, NY, USA, 1355–1364. https://doi.org/10.1145/2702123.2702470

[30] David A Robb, Stefano Padilla, Thomas S Methven, Britta Kalkreuter, and Mike J Chantler. 2017. Image-Based Emotion Feedback: How does the Crowd Feel? And Why?. In DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems. Association for Computing Machinery (ACM), New York, NY, USA, 451–463. https://doi.org/10.1145/3064663.3064665

[31] Peter Gordon Roetzel. 2019. Information Overload in the Information Age: A Review of the Literature from Business Administration, Business Psychology, and Related Disciplines with a Bibliometric Approach and Framework Development. Business Research 12, 2 (7 2019), 479–522. https://doi.org/10.1007/s40685-018-0069-z

[32] Hanna Schneider, Katharina Frison, Julie Wagner, and Andras Butz. 2016. CrowdUX: A Case for Using Widespread and Lightweight Tools in the Quest for UX. In DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems. Association for Computing Machinery (ACM), New York, NY, USA, 415–425. https://doi.org/10.1145/2901790.2901814

[33] Jean Scholtz. 2001. Adaptation of Traditional Usability Testing Methods for Remote Testing. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34). IEEE Computer Society, Los Alamitos, CA, USA, 5030. https://doi.org/10.1109/HICSS.2001.926546

[34] Norbert Seyff, Gregor Ollmann, and Manfred Bortenschlager. 2014. AppEcho: A User-Driven, In Situ Feedback Approach for Mobile Platforms and Applications. In Proceedings of the 1st International Conference on Mobile Software Engineering and Systems. Association for Computing Machinery (ACM), New York, NY, USA, 99–108. https://doi.org/10.1145/2593902.2593927

[35] Melanie Stade, Marc Oriol, Oscar Cabrera, Farnaz Fotrousi, Ronnie Schaniel, Norbert Seyff, and Oleg Schmidt. 2017. Providing a User Forum is not Enough: First Experiences of a Software Company with CrowdRE. In Proceedings - 2017 IEEE 25th International Requirements Engineering Conference Workshops, REW 2017. IEEE, New York, NY, USA, 164–169. https://doi.org/10.1109/REW.2017.21

[36] S Shyam Sundar, Haiyan Jia, T Franklin Waddell, and Yan Huang. 2017. Toward a Theory of Interactive Media Effects (TIME). In The Handbook of the Psychology of Communication Technology. Chichester, West Sussex, UK and Malden, Massachusetts and Boston, Massachusetts, 47–86. https://doi.org/10.1002/9781118426456.ch3

[37] Ralf A.L.F. van Griethuijsen, Michiel W. van Eijck, Helen Haste, Perry J. den Brok, Nigel C. Skinner, Nasser Mansour, Ayse Savran Gencer, and Saouma BouJaoude. 2015. Global Patterns in Students' Views of Science and Interest in Science. Research in Science Education 45, 4 (8 2015), 581–603. https://doi.org/10.1007/s11165-014-9438-6

[38] Karel Vredenburg, Ji Ye Mao, Paul W. Smith, and Tom Carey. 2002. A Survey of User-Centered Design Practice. In Proceedings of the 2002 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 471–478. https://doi.org/10.1145/503376.503460

[39] Helen Wauck, Yu-Chun Yen, Wai-Tat Fu, Elizabeth Gerber, Steven P Dow, and Brian P Bailey. 2017. From in the Class or in the Wild?. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 5580–5591. https://doi.org/10.1145/3025453.3025477

[40] Jane Webster and Hayes Ho. 1997. Audience Engagement in Multimedia Presentations. ACM SIGMIS Database: the DATABASE for Advances in Information Systems 28, 2 (1997), 63–77. https://doi.org/10.1145/264701.264706

[41] 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 2011 CHI Conference Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 143–146. https://doi.org/10.1145/1978942

[42] Y Wayne Wu and Brian P Bailey. 2016. Novices Who Focused or Experts Who Didn't? In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 4086–4097. https://doi.org/10.1145/2858036.2858330

[43] Y Wayne Wu and Brian P Bailey. 2021. Better Feedback from Nicer People: Narrative Empathy and Ingroup Framing Improve Feedback Exchange. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1–20. https://doi.org/10.1145/3432935

[44] Anbang Xu and Brian P Bailey. 2011. A Crowdsourcing Model for Receiving Design Critique. In Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM), New York, NY, USA, 1183–1188. https://doi.org/10.1145/1979742.1979745

[45] Anbang Xu and Brian P Bailey. 2014. A System for Receiving Crowd Feedback on Visual Designs Abstract. In CSCW Companion '14: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported

Cooperative Work & Social Computing. Association for Computing Machinery (ACM), New York, NY, USA, 37–40.

[46] Yu-Chun Grace Yen, Steven P Dow, Elizabeth Gerber, and Brian P Bailey. 2017. Listen to Others, Listen to Yourself. In Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition. Association for Computing Machinery (ACM), New York, NY, USA, 158–170. https://doi.org/10.1145/3059454.3059468

[47] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Björn Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, Vol. 27. Association for Computing Machinery (ACM), New York, NY, USA, 1005–1017. https://doi.org/10.1145/2818048.2819953