Hardhats and Bungaloos: Comparing Crowdsourced Design Feedback with Peer Design Feedback in the Classroom

Jonas Oppenlaender
jonas.oppenlaender@oulu.fi
University of Oulu
Oulu, Finland

Elina Kuosmanen
elina.kuosmanen@oulu.fi
University of Oulu
Oulu, Finland

Andrés Lucero
lucero@acm.org
Aalto University
Espoo, Finland

Simo Hosio
simo.hosio@oulu.fi
University of Oulu
Oulu, Finland

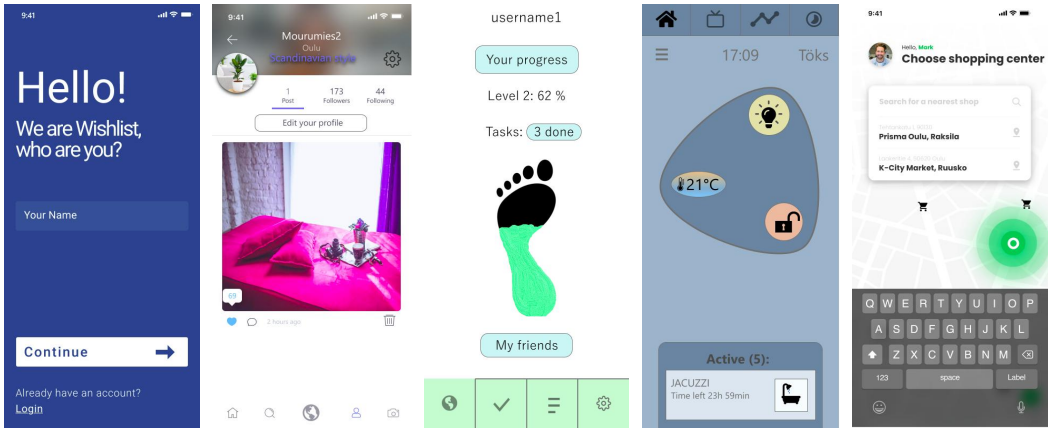


Figure 1: Selected interactive prototypes of mobile applications created by students in the design course.

ABSTRACT

Feedback is an important aspect of design education, and crowdsourcing has emerged as a convenient way to obtain feedback at scale. In this paper, we investigate how crowdsourced design feedback compares to peer design feedback within a design-oriented HCI class and across two metrics: perceived quality and perceived fairness. We also examine the perceived monetary value of crowdsourced feedback, which provides an interesting contrast to the typical requester-centric view of the value of labor on crowdsourcing platforms. Our results reveal that the students (N = 106) perceived the crowdsourced design feedback as inferior to peer design feedback in multiple ways. However, they also identified various positive aspects of the online crowds that peers cannot provide. We discuss the meaning of the findings and provide suggestions for teachers in HCI and other researchers interested in crowd feedback systems on using crowds as a potential complement to peers.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Human-centered computing** → **Empirical studies in HCI**; *Interface design prototyping*; **HCI design and evaluation methods**.

KEYWORDS

crowdsourcing, design feedback, crowd feedback system, classroom study, peer review

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445380

1 INTRODUCTION

Design is an important topic in HCI education. While prior work has established the feasibility of using crowdsourced feedback in design education [6, 17, 37], not many deep investigations into the qualitative expectations and perceptions of students have been documented in the literature.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445380

Our work sets out to replicate and extend prior findings [6, 17, 37], particularly concerning the perceptions of the students. We provide a detailed empirical investigation into how students of a design-oriented undergraduate HCI course perceived and experienced crowdsourced design feedback, and how this feedback compared to peer design feedback in the classroom. In our study, students ($N = 106$) were tasked with brainstorming, designing, and prototyping a functional online interface for a mobile application. Students received formative feedback on the interactive application interface from peers and crowd workers of a paid online crowdsourcing platform (Amazon Mechanical Turk or *MTurk*).

Our study extends prior studies' findings with a comparison of the two sources of feedback across an extensive range of evaluation dimensions related to the perceived quality and felt experience of feedback. We found that the perceived quality of the feedback is shaped by the perceived effectiveness, perceived effort, and the modality of the feedback. Perceived fairness is shaped by agreeableness, valence (i.e., the affective tone and "harshness" of the feedback), diligence, usefulness, and credibility of the feedback. Additionally, we explore how students formulate a monetary valuation of crowdsourced design feedback, which contrasts the typical way of thinking about the monetary value of crowdsourced contributions as rewards paid by the requester. This, we argue, may have implications for crowdsourcing setups in which the teacher as the requester of feedback (and thus the party who is setting the price for the crowdsourced task) is not the receiver and beneficiary of the crowdsourced feedback.

The main contributions of our work are:

(1) An empirical case study on using the crowd from Amazon Mechanical Turk as a source of design feedback in a university-level HCI course.
(2) A detailed analysis of the perceived attributes of formative design feedback from two sources: peers and the crowd from Amazon Mechanical Turk.
(3) An exploration of how students formulate a monetary valuation of crowdsourced design feedback.
(4) A discussion on the implications and how teachers might best leverage crowdsourced design feedback in their courses.

Overall, the students in our study found the formative design feedback from peers to be of higher quality across all of the analyzed criteria, except for valence. However, the diversity of people from different backgrounds was mentioned as a particularly positive aspect of the crowdsourced feedback, and approximately 25% of the participants preferred MTurk feedback over peer feedback. Concerning the monetary value of crowdsourced feedback, participant responses differed drastically from the pay that crowd workers typically receive on MTurk. Our findings are informative to researchers working on crowd feedback systems as well as teachers of design-oriented HCI courses who wish to explore crowdsourced feedback as a way to conveniently expose students to feedback from outside the course itself.

2 RELATED WORK

The field of HCI has always had an interest in learning and education, for example by establishing a new dedicated CHI subcommittee on "Learning, Education, and Families" and a Special Interest Group (SIG) at CHI '19 [30]. Our work touches on design feedback, feedback collection via crowdsourcing, and the application of crowdsourced feedback in the classroom.

2.1 Design Feedback

Feedback is an important component of design education. Feedback communicates a notion of a standard to the learner, and learners typically strive to minimize the feedback-standard discrepancy [21]. The overall goal of design education is to raise the learner's conception of standard to that of the teacher [34]. One important mechanism to realize this goal is design feedback.

Design feedback can be given at four different levels (product, process, self-regulation, and self [13]), and the feedback can be formative or summative. While summative feedback is geared towards grading results at the end of a course, formative feedback guides students in improving their work [34]. According to the theory of formative assessment [34], formative feedback is instrumental to developing expertise. Formative peer feedback for in-progress work may, for instance, improve course outcomes [23]. Effective feedback provides specific information about a learner's current performance together with explanations and concrete examples [1]. Successful feedback is specific, critical, and actionable [34, 44].

The term *critique* is sometimes used as synonym for design feedback in related literature (e.g., in [27]). And indeed, a critique is "the communication of a reasoned opinion about an artifact or a design" [9]. Traditionally, however, a (studio) critique refers to a formal, co-located feedback setting in which knowledgeable peers or experts provide feedback to student learners [4]. Critiques are a common practice in design education as a form of assessment. In a critique, co-located students discuss and evaluate their sketches, collages, or designs. Critiques are also practiced among design colleagues as a means to receive feedback.

In this work, we concentrate on product-based formative feedback for a design artifact. We follow a process-agnostic conceptualization of feedback as information provided by a *feedback provider* (a teacher, peer, or other kind of *reviewer* of a work) to the *feedback receiver* (in our case, a university-level student) [13].

2.2 Crowdsourcing Design Feedback

Peer feedback has become a common pedagogical tool especially in online teaching and when the instructors are not always available [7, 24]. Crowdsourcing offers an even more scalable approach to feedback. Researchers have in the past looked into the trade-offs of these approaches. Nguyen et al. found that design feedback from anonymous sources was perceived more positively compared to feedback from peers or an authority [27]. Anonymous feedback providers may give feedback that contains more specific criticism and praise, and thus be rated as more useful by the feedback receiver [17]. Anonymity may further help minimize power differences between the feedback provider and the feedback receiver [4]. To this end, crowdsourcing platforms, such as Amazon Mechanical Turk (*MTurk*), may provide a fruitful ground for eliciting design feedback from a large and diverse group of anonymous people. On MTurk, requesters (the feedback receivers) publish short "Human Intelligence Tasks" (*HITs*) for anonymous workers (the

feedback providers) to complete in exchange for a small monetary reward [16].

The microtask crowdsourcing model is, however, – by its nature – troubled with a number of issues that may negatively affect the provision of feedback. Survey satisficing may inflate reliability and validity of collected data [11] and could negatively impact the quantity and quality of the collected feedback. The power differences between workers and requesters manifest in low payment to workers [12]. Workers may further be subject to numerous biases, such as the observer effect. On the other hand, feedback from peers may be positively biased [36], and students were shown to appreciate and prefer feedback from external and anonymous sources [6, 27].

Researchers created a number of web-based systems to explore and investigate the potential of MTurk for collecting feedback on graphic designs. *Voyant* [41] is a system for eliciting perception-based feedback on graphic designs from a non-expert crowd. The system aims to capture the first impressions and how well a graphic design meets its stated goals and design guidelines. *CrowdUI* [29] is a system for eliciting and aggregating visual feedback on the design of a website. The system enables users to modify a website's user interface. The user-generated design suggestions are presented to the feedback requester in aggregated form. *CrowdCrit* [25] gathers formative feedback on graphic designs from crowd workers and clients. Designers found the feedback provided by the system to be helpful and appreciated the specificity and level of detail of the feedback. Feedback from crowd workers was however found to be "more generic." In *Paragon* [19], feedback providers complement written feedback on graphic designs with visual design examples. Feedback provided in this manner was found to be more novel, specific, and actionable. *SIMPLEX* [28] is a system to gather summative feedback for designs and artworks from a situated crowd using two public displays. *ZIPT* [5] is a system that – unlike the above systems that elicit feedback on static graphic designs – uses virtualization technology to enable a crowd to conduct remote user tests on mobile applications. This setup is similar to the one used in our study, except that feedback providers in our case provided feedback on a high-fidelity web-based prototype, not a real mobile application.

Researchers further investigated strategies for improving the quality of crowdsourced design feedback. Rubrics are an effective way to structure feedback and raise its perceived value, valence (i.e., affective tone), and specificity [44]. Scaffolds were shown to support students in reflecting on the feedback received [2]. Structured workflows were found to generate feedback that was more interpretative, diverse and critical than free-form prompts [42]. Further, guiding questions formulated by the feedback receiver were found to be effective scaffolds to facilitate the exchange of feedback among peers [3]. Measures such as the above have been found to improve the quality of crowdsourced design feedback.

In contrast to the above quality measures, our study provides insights into the felt experience of evaluating feedback. The participants in our study were given the responses from Amazon Mechanical Turk *as is*, without any filtering of responses or other post-hoc measures for improving data quality. We deliberately chose to compare the "unfiltered" feedback from the two sources. We decided against filtering the crowdsourced data as we believe this is the fairest way of comparing the two feedback sources without paying favors to one source. This decision avoids introducing bias to the study and best reflects what can be expected from the two types of feedback providers. However, we did employ standard qualification measures widely used in academic studies on MTurk.

2.3 Crowdsourced Feedback in the Classroom

Dow et al. established the feasibility of using crowds from MTurk for collecting design feedback in the classroom [6]. In Dow et al.'s studies, the crowd provided feedback along four key stages of the innovation process (needfinding, ideating, testing, and pitching). Students found feedback from crowd workers to be beneficial in all four stages and the crowdsourced feedback helped students ground their design efforts in real-world opportunities. Most relevant to our study, testing early-stage storyboards with the online crowd was found to help students uncover issues and directions for future work.

Xu et al. used a crowd feedback system to provide 10 students in a visual design course with formative feedback from workers on MTurk [42]. The authors found that formative feedback from the crowd prompted students to change and improve their designs. Compared with the feedback from experts, the crowdsourced feedback was however not found to agree on whether a design met its communicative goals.

In the work by Wauck et al., students created prototypes of user interfaces in a project-based design course. Students received feedback from their peers and three different online crowds: the students' own social network, online communities, and Amazon Mechanical Turk. The authors measured the quality, quantity, and valence of the feedback, and how it was acted upon. Student peers were found to provide feedback that was higher in perceived quality, more acted upon, and longer than feedback from online communities. Further, summative feedback from both peers and online communities was found to be more negative in valence than formative feedback at an early stage in the design project.

Our study is similar to the study by Wauck et al., both in its focus and study design. Our study is therefore partially a replication study of the work by Wauck et al.. Our contribution to HCI is the confirmation of the prior study's findings [14]. The study by Wauck et al. answers three research questions, only one of which is student perceptions, whereas our work focuses only on the student perceptions, but extends the prior study to more dimensions qualitatively and quantitatively. Our work therefore extends and contributes a clear increment to prior studies of feedback in the classroom with an in-depth analysis of students perceptions.

Prior literature found that student designers may attribute value to feedback along a number of different criteria, such as the feedback's *quantity* [25, 37, 44], *specificity* [2, 3, 17, 19, 25, 34, 44], *criticality* [2, 3, 34, 44], *valence*, *affect*, or *sentiment* [3, 27, 37, 42, 44], *helpfulness* [5, 22, 25, 41], *fairness* [44], and *actionability* [2, 3, 19, 34, 37, 44]. Students may, for instance, find specific and emotionally positive feedback useful and students are likely to appreciate longer and more actionable feedback with clear justifications [44]. In our study, we provide a mixed-method investigation into these aspects of formative design feedback in the context of a university-level design course.

3 THE STUDY

Our aim was to investigate how students perceive and experience crowdsourced formative feedback in the classroom and how this feedback compares to peer feedback from the students' classmates. The study was conducted in the context of a design course at the University of Oulu, Finland.

3.1 HCI Course and Design Projects

The study was conducted in a 9-week undergraduate course (521145A "Human Computer Interaction") between October and December 2019. The course followed a project-based learning approach. In the beginning of the course, students formed 42 groups with three members each around self-chosen topics. Each group of students was tasked to brainstorm, design, and prototype a user interface of a mobile application. Each group first sketched their idea with pen and paper and evaluated this paper prototype in a user study with their classmates. The groups then transferred their designs into a digital prototype using *Adobe XD* (a user interface design tool), while taking into account the lessons-learned from the user studies. The interactive prototypes were published online using Adobe XD's functionalities, which allowed us to invite both students and workers from a crowdsourcing platform to operate and review the interactive application prototypes. Each group received formative feedback from other students and crowd workers. After reviewing the formative feedback, the groups acted on the feedback received and submitted a revised version of the digital prototype for grading. Selected examples of the interactive designs created during the course are depicted in Figure 1.

3.2 Study Design

Our study design choices were shaped by practical and pedagogical considerations and informed by our study being a live, authentic teaching exercise rather than a controlled laboratory experiment on feedback quality. Our study therefore needs to be viewed in light of being a case study of feedback with ecological validity in the classroom.

We conducted a within-subject study in the context of the HCI course. Each of the 42 groups of students was provided a zip file (the "feedback package") with formative feedback from two sources: nine of their peers and nine crowd workers. The number of feedback items was determined based on the group size of the students. Since each group consisted of three members, we asked each student to review the work of three other teams. Students were briefly introduced to the notion of crowdsourcing in the course lectures. The within-subject design was mandated by the university's teaching rules, to provide each group of students a similar learning experience and not to disadvantage some groups in the graded course. Students were individually asked to comment on the formative feedback in a survey (the "final questionnaire") at the end of the course. In the following section, we describe the procedure for collecting feedback, how feedback was provided to the students, and the final questionnaire.

3.2.1 Apparatus and materials. Feedback was collected using the web-based survey instrument by Hosio et al. [15] (depicted in Figure 2). The interface included 14 questions, of which 13 were elicited with sliders. Both students and crowd workers (henceforth: the *reviewers*) received the same instructions and provided feedback with the same interface, with one minor difference. Students were presented with a list of all groups and navigated to the group they were assigned to review. For the review by crowd workers, the sidebar with group names was hidden to not distract the workers. Instead, the task was set up so that the worker would only see the one design that was to be rated during that task. For the students, providing peer feedback was a fixed part of the coursework, and students provided feedback individually in their self-study hours outside the scheduled exercises.

Using the web-based interface, the reviewers rated the mobile application prototype on several criteria (each on a 7-point anchored Likert scale; see Figure 2). We elicited the novelty and usefulness of the application prototype as two important components of product-based creativity [33]. We further elicited practicality and a rating of the design. Each criteria was explained with a short question. For instance, novelty was explained with "Has the application been thought of by others?" Reviewers were asked to rate their success in following the set of tasks provided by the student group. Further, the interface included the 8-item version of the User Experience Questionnaire (UEQ-S) [35]. The UEQ-S measures several aspects of user experience on dichotomous scales, such as "Boring-Exciting," "Confusing-Clear," and "Complicated-Easy." In the last item, reviewers were asked to provide an exhaustive open-ended justification for how they scored the different criteria.

Feedback Package. The collected feedback from the reviewers was bundled in a zip file (see Table 1). Each feedback package contained raw data collected with the web-based survey instrument, a graphical summary of Likert-scale responses (split between peer and crowd feedback), and two text files with the open-ended feedback from nine student reviewers and nine crowd workers, respectively. The feedback package was accompanied with detailed instructions for the students. A full example of a feedback package is available in the auxiliary material.

3.2.2 Study procedure. Students individually inspected the feedback package at the end of the course (in their self-study hours outside of the scheduled exercises). Feedback was not given in a specific order, and the students could freely choose which one to consume first. The students then evaluated the contents of the feedback package in an online questionnaire. All related ethical procedures were followed as required by our University. Students were asked to consent to their data being used for the purpose of the academic study prior to completing the study. Students were informed that declining consent did not affect their course grades. Students were specifically instructed to fill out the final questionnaire individually, not in teams. As an incentive, 5 points counting towards the course credit were awarded to the students who calculated the three components of the UEQ-S (Overall, Pragmatic, and Hedonic Quality) from the raw data.

The final questionnaire consisted of 18 items including demographic questions, an evaluation of the formative feedback, and the satisfaction with the course. The questionnaire items related to the evaluation of the feedback are listed in tables 2 and 3. The full final questionnaire can be found in the auxiliary material.

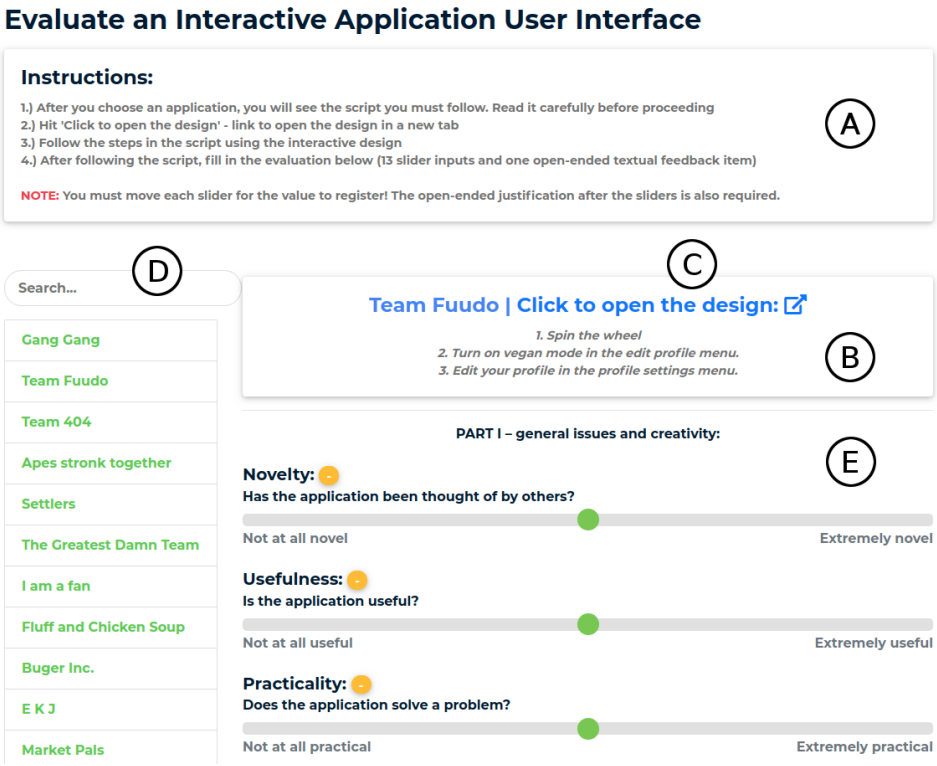


Figure 2: Web-based interface for feedback collection. A set of general instructions was provided to the reviewers (A). Each student group provided their own set of tasks to be completed in the mobile application (B). A click on a link (C) directed the reviewer to the interactive prototype of the mobile application hosted on Adobe XD resource servers. The reviewer was instructed to rate the mobile application along a number of dimensions (E). Students could access the groups they were assigned to review in the sidebar (D). A screenshot of the full interface is available in the auxiliary material.

Table 1: Feedback package with peer and crowdsourced feedback provided to the students.

File	`data.xlsx`	`summary.png`	`summary_openended_class.txt`	`summary_openended_mturk.txt`
Description	Quantitative responses (raw data) collected with the survey instrument	Graphical summary (bar chart) of the feedback, split between peer and MTurk feedback	Open-ended feedback from student peers	Open-ended feedback from workers on MTurk

The questionnaire elicited detailed justifications for the students' preference of feedback (see Table 3). We first inquired which of the two feedback sources the students preferred overall. This first open-ended item was given to capture students' reasoning in their own terms, without imposing a structure or limiting the response. The following three items were structured around three principles of good feedback: *effectiveness*, *fairness*, *actionability* [34, 44]. Students were next asked to quantitatively rate the feedback on 7-point Likert scales (see Table 2). Effectiveness was measured with items related to effective feedback [1]: *specificity*, *actionability*, and *explanations*. *Fairness* and *actionability* had their own items, and in addition, students were asked to judge the overall *valence* of each feedback source, the *relevance* of the feedback to the application prototype, the overall *quality* of the feedback, and the *satisfaction* with the respective feedback source. Finally, students were asked to provide their personal estimate of the *monetary value* of the feedback provided by crowd workers (*"Looking only at the feedback from online workers on Mechanical Turk for your own group, how much do you think this feedback is worth, in money (Euros)?"*). Students were specifically instructed to estimate the cost of *"all of the feedback from MTurk combined (but for your own group) – not for an individual feedback item among the many."*

3.3 Participants

3.3.1 Crowd workers. We invited crowd workers from Amazon Mechanical Turk to review the 42 mobile application prototypes created by the student groups. Each prototype was assigned to nine different crowd workers. Workers were recruited following best

practices (HIT approval rate greater than 95% and number of HITs approved greater than 100). Similar qualification criteria are widely used in academic studies [31]. In total, 387 HITs (with an attrition rate of 2.3%) were completed by 328 unique workers. Workers were paid US$ 1 per HIT completed.

3.3.2 Students. Out of 112 students who submitted the final questionnaire, 106 students (P1–P106, aged from 18 to 58, $M = 23.3$ years, $SD = 5.6$ years) consented to contribute their data to the study. As the course was offered by a Faculty of Technology, the gender distribution of the participants reflects the bias that is often found in technology-dominated subject areas (82 male, 20 female, 1 non-binary/third gender, and 3 undisclosed). The course was primarily aimed at Bachelor students (78 participants), but Master students were allowed to attend (28 participants). Over half the students (62 participants) were enrolled in a degree program in Computer Science. However, the course was also attended by students from other disciplines, such as Industrial Engineering (24 participants), Electrical Engineering (12 participants), English Philology (1 participant), Medicine (1 participant), and other degree programs (6 participants).

3.4 Methodology

We focus our mixed-method analysis on the students' evaluation of the feedback in the final questionnaire. We only briefly report on the feedback itself, because we want to foreground the subjective experience of formative design feedback in this paper.

We compared different criteria of how students perceived the feedback, as described in Section 3.2.2. Since the data was not following a normal distribution (according to Shapiro-Wilk's test), we used paired Wilcoxon rank sum tests to evaluate the differences in the students' perception of the peer feedback and the crowdsourced feedback. Besides investigating the students' perspective on the feedback, we quantitatively analyzed and compared the open-ended feedback along three criteria contributing to the quality of the feedback: the *length* of the feedback (measured by the number of characters in the feedback item), the amount of noise in the feedback (as measured by the presence of unrelated and nonsensical words), and the *effort* spent on providing the feedback (as measured by the time taken to complete both the Likert-scale feedback and the open-ended feedback).

Qualitatively, the three open-ended items (see Table 3) were analyzed following the guidelines for content analysis [38]. One researcher first familiarized himself with the data and extracted verbatim terms from the first item of the questionnaire in which students provided reasons for their overall preference of feedback. The researcher consolidated the extracted terms into codes, while also considering the responses to the other two open-ended questionnaire items about the effectiveness and fairness of the feedback (Table 3). Each code was categorized on whether it was a positive or negative statement about peer or crowd feedback, respectively. Subsequently, the researcher iteratively and inductively grouped the codes into themes and sub-themes. Next, a codebook was developed with descriptions for each code. After discussing the codebook, the first and second authors of this paper and one additional student individually coded all responses to the first open-ended item of the questionnaire. Inter-rater reliability improved after reconciling differences in two discussions and adjusting the codebook. The final inter-rater reliability among the three raters (as measured by Fleiss Kappa [10]) was $\kappa = 0.83$.

4 QUANTITATIVE COMPARISON OF PEER AND CROWDSOURCED FEEDBACK

We found significant differences in how the students perceived the feedback from peers and crowd workers in all criteria except valence (see Table 2 on page 7).

Peer feedback was significantly more *specific*, more *actionable*, and contained more *explanations* than the crowdsourced feedback in the opinion of the students (each $p < 0.001$). Peer feedback was also deemed more *relevant* to the mobile application prototype and of better *quality* than the crowdsourced feedback (each $p < 0.001$). Overall, students were more *satisfied* with the peer feedback compared to the crowdsourced feedback from MTurk ($p < 0.001$). On average, the feedback from MTurk was perceived slightly more positive in affective tone than the peer feedback ($M = 4.51$ for crowdsourced feedback versus $M = 4.49$ for peer feedback). However, no statistical difference between peer feedback and crowdsourced feedback was found in terms of *valence* ($p > 0.05$).

An inspection of the feedback revealed that some of the answers to the open-ended item were extremely short and contained either non-existing words (e.g., *"uehi," "asdf,"* and *"adse"*), only numbers, or other words not relevant to the task (such as *"yes," "no," "this app,"* and *"NICE"*). We manually curated a list of words and found the amount of noise greatly differed between the two feedback sources. Only 0.3% ($N = 1$) of the open-ended peer feedback contained noise, compared to 12.2% ($N = 47$) of the feedback from MTurk.

Related to the amount of noise, the peer feedback contained significantly longer explanations compared to the crowdsourced feedback (peer feedback $M = 296$ characters, $SD = 287$ characters versus crowdsourced feedback $M = 117$ characters, $SD = 173$ characters, $p < 0.001$, $W = 105454.5$). The amount of answers with less than 10 characters (including the noise) was 0.6% in the peer feedback and 23.4% in the crowdsourced feedback.

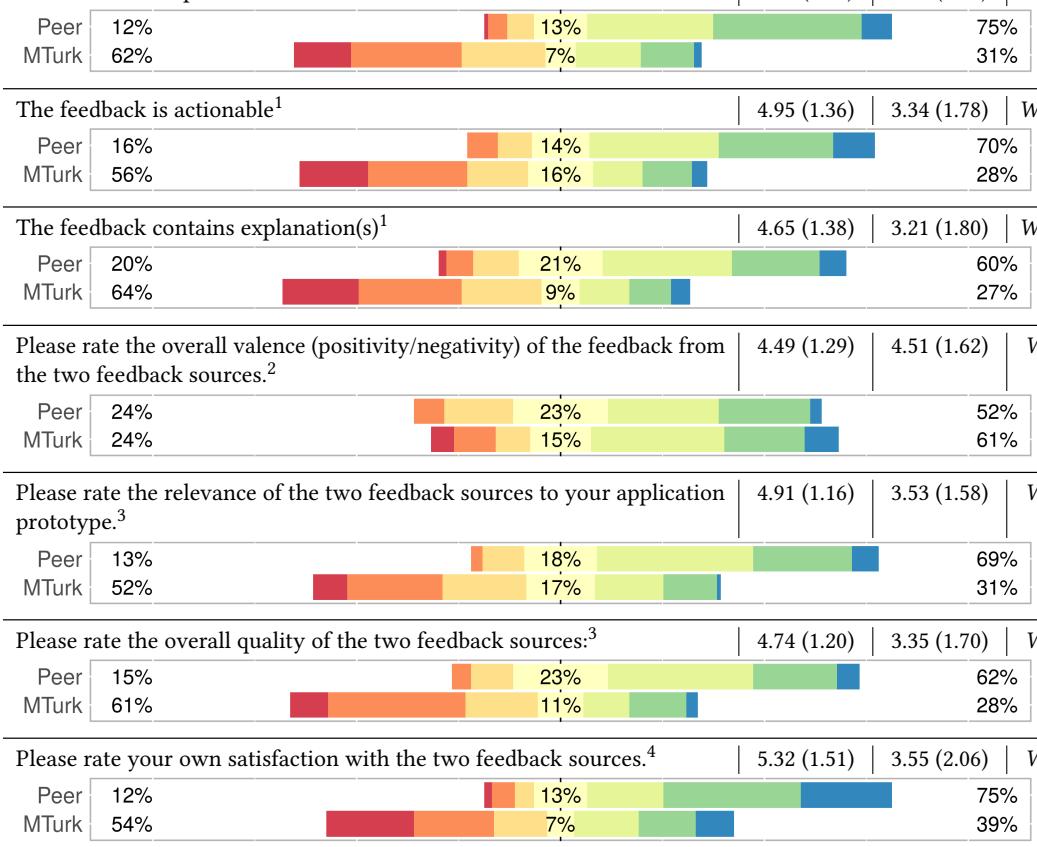
The time spent on writing the feedback varied between 18.2 seconds and 22 hours. As the extremely long times do not reflect the actual time spent on the task, we removed times longer than one hour in the following comparison. One hour was the the maximum time available for a crowd worker to complete the survey in our data collection campaign on MTurk. There were 22 observations over 1 hour in the peer feedback, and none in the crowdsourced feedback. On average, the class took more time to provide feedback ($M = 12.2$ min, $Min = 18.3$ seconds, $Max = 54.2$ min, $SD = 10.2$ min) than the crowd workers ($M = 4.6$ min, $Min = 18.2$ seconds, $Max = 42.2$ min, $SD = 5.2$ min; $p < 0.001$, $W = 103461.5$).

5 QUALITATIVE FINDINGS

We coded the open-ended questions to determine the preference of each student for either peer or crowdsourced feedback (see Table 3). The majority of the students ($N = 79$; 74.5%) voiced a preference for peer feedback, but a not entirely insignificant minority ($N = 26$; 24.5%) preferred the feedback from MTurk. Overall, students found the peer feedback more *useful*, more *effective*, more *actionable*, and of greater *fairness*. In contrast, students often noticed and complained about the low perceived *effort* in the crowdsourced feedback. We elaborate on each of these aspects in the following sections.

Table 2: Likert-scale items in the final questionnaire given to students at the end of the course. The Likert plots reflect the percentage of participants who agreed, disagreed, or were neutral to a given statement. For example, 62% of the students thought the formative feedback from MTurk workers was not specific, 7% neither agreed nor disagreed, and 31% agreed to this statement.

Feedback item	Peer feedback M (SD)	Crowd feedback M (SD)	Wilcoxon rank sum
The feedback is specific[1]	5.08 (1.28)	3.31 (1.72)	$W = 3897.5$
The feedback is actionable[1]	4.95 (1.36)	3.34 (1.78)	$W = 3482$
The feedback contains explanation(s)[1]	4.65 (1.38)	3.21 (1.80)	$W = 3542$
Please rate the overall valence (positivity/negativity) of the feedback from the two feedback sources.[2]	4.49 (1.29)	4.51 (1.62)	$W = 2036.5$
Please rate the relevance of the two feedback sources to your application prototype.[3]	4.91 (1.16)	3.53 (1.58)	$W = 3505.5$
Please rate the overall quality of the two feedback sources:[3]	4.74 (1.20)	3.35 (1.70)	$W = 8435.5$
Please rate your own satisfaction with the two feedback sources.[4]	5.32 (1.51)	3.55 (2.06)	$W = 8480$



Response 1 2 3 4 5 6 7

[1] 1 – Strongly Disagree ... 7 – Strongly Agree
[2] 1 – Strongly Negative ... 7 – Strongly Positive
[3] 1 – Very poor, 2 – Poor, 3 – Fair, 4 – Good, 5 – Very good, 6 – Excellent, 7 – Exceptional
[4] 1 – Very dissatisfied, 2 – Moderately dissatisfied, 3 – Slightly dissatisfied, 4 – Neutral, 5 – Slightly satisfied, 6 – Moderately satisfied, 7 – Very satisfied
*** $p < 0.001$

We structure our qualitative findings around three main themes of experiencing feedback: *feedback quality* (with sub-themes *effectiveness*, *effort*, and *modality*), *feedback fairness* (with sub-themes *agreeableness*, *valence*, *diligence*, *usefulness*, and *credibility*), and the *monetary valuation* of the feedback.

5.1 Perceived Quality of the Feedback

5.1.1 Perceived effectiveness of the feedback. The overall perception of the effectiveness of the feedback (from both sources) ranged from *"not that effective"* (P3) to *"highly effective"* (P5).

Peer feedback was overall perceived as more meaningful and comprehensive. Peers provided more constructive and actionable suggestions for improvement than the crowdsourced feedback. Several students lauded the peer feedback for *"pinpointing design flaws"* (P18) and specific elements to improve, such as buttons, font sizes, wording, and other *"little things that bugged [the reviewers]"* (P5).

The open-ended crowdsourced feedback, on the other hand, was seen as being more *"general"* (e.g., P7) and, as a consequence, less useful and actionable by the majority of students. Several students complained about the low number of takeaways found in the crowdsourced feedback.

Table 3: Qualitative items of the final questionnaire given to students at the end of the course, with coded preference for peer feedback and crowd feedback. For example, 79 students overall preferred the peer feedback, 26 the crowdsourced feedback, and one student could not decide.

Item	Type	Peer	MTurk	undecided	n/a[1]
Which of the two feedback sources (in-class from other students or crowdsourcing from online workers) do you **prefer**, and why?	Open-ended	79	26	1	–
How **effective** did you find the feedback (from both sources combined) in helping you understand what you did well and how you could do better?	Open-ended	–	–	–	–
Which of the two feedback sources (in-class or crowdsourcing) do you think was more **fair**, and why?	Open-ended	62	34	9	1
Which of the two feedback sources (in-class or crowdsourcing) contained items that you would be able to act on – that is, which feedback is more **actionable**?	Multiple-choice[2]	84	22	–	–

[1] Preference could not be determined from the response.
[2] "In-class feedback from other students" or "MTurk feedback from online workers"

Only few students ($N = 9$) found both sources equally effective. P8, for instance, thought that *"the class noticed the bad sides and the online workers noticed the good sides of the app."* P21 searched the fault in his own group, as he thought that *"the design was not realistic and attractive for a general public, and some features were not that clear"*, but expressed optimism that *"after [implementing the review] we think that the opinions would change."*

5.1.2 Perceived effort. Students noticed a difference in effort put into writing the open-ended feedback. The peer feedback was perceived as being more elaborate. The length of the feedback affected the perception of usefulness, as students were able to find fewer takeaways in the crowdsourced feedback. Crowdsourced feedback was often perceived as nonsensical, primarily due to some of the responses being extremely short and unrelated to the task: *"My group didn't learn anything useful from online workers open ended questions, it had answers like "24", "god like app" and "Useful app :)". All of our classmates answered with many rows of text"* (P16).

Students recognized that their peers had put more thought and time into writing the feedback, whereas the *"online workers just tried to speedrun the questions"* (P91). Two students suspected the feedback by MTurk workers was provided by *"bots and other ESLs who probably didn't even give it anything but a glance to get their pennies"* (P3), and that *"the slider selections were done randomly by these bots"* (P1).

5.1.3 Modality of the feedback. Students generally preferred the open-ended feedback over the raw data and summary chart. The written justifications allowed the students to identify specific problems in their user interfaces and areas to improve. In contrast, the numerical raw data and the summary chart was perceived as less helpful, as exemplified by the comment from P6 who thought *"the numerical scores weren't really helpful; too often a low score wasn't justified in the open-ended feedback, and it was hard to understand what we could improve."* The numerical feedback was only *"useful for finding the most obvious flaws and getting an overall feeling of how the people reacted"* (P27). Nevertheless, students appreciated the summary chart, as they *"found it useful to have two different groups of people to evaluate our application [...] It forced us to think how we could make it more appealing to both groups"* (P43).

When evaluating the numerical feedback, students primarily looked for differences between the two sources of feedback in the summary chart, not commonalities. "Variety" was mentioned several times as criteria to look for in the summary chart. P92, for instance, mentioned that *"in the [Likert-scale] questions there was not enough variety to see which areas were good and which needed improvement."* Students liked to contrast the feedback from the two sources, as it allowed them to identify the weaknesses and strengths in their designs. For instance, P10 mentioned that *"since both sources are not correlated, it was easy to identify the main design failures in the app and prioritize and solve them."*

5.2 Perceived Fairness of the Feedback

Approximately two thirds of the students (62 students) found the peer feedback was more fair than crowdsourced feedback, compared to 34 students who thought the crowdsourced feedback was more fair (see Table 3). Students primarily perceived the feedback's fairness along five dimensions: (a) *agreeableness*, (b) *valence*, (c) *diligence*, (d) *usefulness*, and (e) *credibility*. The five dimensions were mentioned by approximately the same number of students.

5.2.1 Agreeableness. The students' perception of fairness was strongly determined by how agreeable they perceived the feedback to be. Sensible and well justified feedback was easier to reconcile with their own view in this regard. Constructive and well-thought out feedback was overall perceived to be more fair than feedback lacking these attributes. Peer feedback was generally perceived to possess more of the above attributes than crowdsourced feedback.

Responses from students who preferred crowdsourced design feedback in terms of fairness were often motivated by the lack of criticism which made the crowdsourced feedback "more agreeable," such as the comment from P16 who found crowdsourced feedback was *"much more positive and [workers] seemed to like our app. Fellow students were much more critical. I think our class tried to be too critical and find every little thing that was wrong."* Fairness, in the minority group who thought crowdsourced feedback was more fair,

was simply a matter of the crowd finding *"less problems with the application"* (P5).

Several students perceived feedback with a variety of different viewpoints as more fair. For these students, peer feedback contained a greater diversity of viewpoints and was thus judged to be more fair than crowdsourced feedback.

5.2.2 Perceived valence. As for the contribution of the affective tone (i.e., valence of the feedback) to the students' sense of fairness, students perceived the peer feedback as more critical, and much more harsh and negative in tone than the crowdsourced feedback. The criticism and negativity divided the students. While the majority of students appreciated critical feedback and perceived it as fair, a minority perceived the critical peer feedback as less fair than crowdsourced feedback.

Among the students who thought that critical peer feedback was fair, the crowdsourced feedback was often found to be overly positive and not critical enough to be useful. P1, for instance, found peer feedback more fair because *"even the little feedback we got from [MTurk] was way too positive and approving and did not criticize the obvious problems the prototype had."* Similarly, P8 stated that he preferred peer feedback, because peers had *"given more negative feedback compared to Mturk people and some of those comments helped lot to further improve and redesign our app."*

Feedback from MTurk was generally perceived to be more approving and praiseful. However, not all students considered positive feedback as being correlated with fairness. Of the 16 students who mentioned MTurk feedback being positive, six found the crowdsourced feedback was fair, but 10 thought the crowdsourced feedback was unfair because of its overly positive praise. As for the reasons of why peer feedback was more critical, P92 speculated that *"working with our own designs during the course made us students more demanding towards each other's designs than the MTurk workers."*

5.2.3 Perceived diligence. A high quantity of feedback positively contributed to the students' sense of fairness. The feedback from MTurk lacked in this regard and was perceived as less fair than the peer feedback due to the often short replies. In the same vein, some students judged the MTurk feedback to be less fair because of its superficiality, lower clarity and lower specificity, as exemplified by the comments from P85 who thought that *"peer feedback was more fair because the some crowdsourced feedback was not clear,"* and P8, who thought that *"Mturk people gave more general comments which were useful too, while peers were more specific on errors and mistakes – like spellings."* Internal consistency and absence of contradictions was mentioned by a few students as contributing positively to the fairness of the feedback.

5.2.4 Perceived usefulness. Feedback with a high number of suggestions and helpful hints was perceived more fair than feedback lacking these features. The effectiveness of the feedback and its perceived value also contributed to the students' sense of fairness. Peer feedback overall was perceived more useful and fair than crowdsourced feedback. One contributing factor to this sense of fairness was the context of the feedback's inception. The peers were familiar with the design task and had all undergone the same training. The peer feedback was thus perceived as more useful, as exemplified by the statement from P28:

> *"the class seemed to know the context better, and gave more relevant feedback. Much of the MTurk feedback was useless."*

On the other hand, some students thought the crowdsourced feedback was more fair because of the crowd having an outsider's perspective on the application. P88, for instance, thought that *"crowdsourcing was fairer because it felt like there was no obligation for the testers to be nice."* The external reviewer having an impartial point of view also contributed to the students' perception of the credibility of the feedback, as detailed in the next section.

5.2.5 Perceived credibility. The credibility of the feedback provider was an especially strong argument among the students who preferred the feedback from MTurk in terms of fairness. Among these students, the feedback from external workers was perceived to be more objective than the feedback from classmates. The *"external and impartial point of view"* (P19) was appreciated by several students and contributed to the sense of credibility. The workers were neither familiar with the course nor the students, and were thus perceived as being less biased, which positively contributed to the trust into the workers' feedback.

The majority of students, however, perceived peer feedback as more credible than crowdsourced feedback. Among this majority, peer feedback was perceived as more serious and sincere, more realistic, and better scoped to the course. Feedback from MTurk, on the other hand, was described as containing *"a lot of joke answers"* (P15). Some of the distrust into the crowdsourced feedback was, however, motivated by the students being unfamiliar with the external source of feedback. P47, for example, commented that he preferred *"feedback from other students because I am not familiar with MTurk, so it is difficult to estimate the reliability of it."*

5.3 Estimating the Monetary Value of Crowdsourced Feedback

We asked each student to quantify the monetary value of the crowdsourced feedback for the group work (see Section 3.2.2). In this section, we provide an analysis of the responses to this item.

5.3.1 Monetary value. We extracted the monetary value from each response (where possible and using the mid-point of a range in seven of the 106 responses). Unrealistically high values (e.g., 5000 EUR) were removed using the upper 1.5 interquartile range (IQR) as a cut-off. After the removal of outliers, the monetary value of the crowdsourced feedback ranged from 0 EUR to 25 EUR, with a mean value of 6.50 EUR ($Median$ = 5 EUR, $Q1$ = 0.80 EUR, $Q3$ = 10 EUR).

5.3.2 Factors affecting the valuation. Twenty-nine of the 106 students provided written explanations for their choice of monetary value. In the remainder of this section, we provide an account of the 29 students' thoughts with specific attention to how students voiced and reasoned about their valuation of the feedback from crowd workers. The overall sentiment of the feedback's value ranged from positive to extremely negative. Two main criteria for valuing the crowdsourced feedback emerged from the students' responses: perceived usefulness and professionalism (see Table 4). The valuation

of the feedback fell in two camps: students who thought the feedback was completely worthless (we refer to them as *"hardhats"* due to their strict, matter-of-factly and engineer-like approach to estimating the feedback's value), and students who attributed some value to the feedback (the *"bungaloos"* due to their bona fide way of valuing the feedback driven by conflated expectations).

Among the hardhats, the bulk of the comments mentioned not wanting to pay or not valuing the feedback at all. This was often justified with the low quality of the feedback. P13, for example, would not pay anything *"because the feedback was extremely poor."* Other students in this camp expressed their sense of value in more drastic ways, often motivated by the low usefulness and low relevancy of the crowdsourced feedback for the design task. P48, for instance, thought that *"Mturk should pay us to read that rubbish, there were only two useful feedbacks and for them maybe few euros."* Similarly, P28 complained that *"only one comment was actually relevant"*, and *"even if they paid you, this feedback would not be worth anything."* While the potential of the crowdsourced feedback was recognized, some students were disappointed, thinking that the feedback was worth *"absolutely nothing, it was a disgrace and a waste of money for at least most of our feedback. I've seen some of our classmates in other groups getting fair feedback and specific ones but ours was just...blatantly saying extremely disappointing"* (P4). P98 expressed regret that the crowdsourced feedback was worth *"sadly nothing because it literally gave us nothing to work with."*

A perceived low specificity of the crowdsourced feedback was a common theme in the responses from both camps, as exemplified by the comment of P89 (hardhat) who thought that some workers *"talked about the idea and not the design, and to top it all lot of the critics were not specific at all and so short we couldn't take anything from them (like 'GOOD')."* A suspicion of low effort affected the students' valuation of the feedback, as highlighted by P99 (bungaloo) who thought that some workers *"didn't even spend 1 minute on filling the feedback form so those deserve 1€ max. We got a very detailed review also though which was worth 3-5€ imo."*

Only one of the 29 students explicitly mentioned considering the minimum wage of workers on MTurk. This student (P12, bungaloo) placed the value of the feedback *"somewhere between whatever the minimum fee on MTurk was (90 cent or so I don't remember exactly) to few euros at max (maybe 5-6 if generous)."*

In the camp of bungaloos, the low amount of information in the feedback was recognized by several students, with also many complaints about the relevance and usefulness of the feedback. P26, for instance, complained that *"some of the feedback had basically no content in them"* and *"it can't be more than maybe 30€ because if it's more than that, the whole thing is a rip off."* Similarly, P43 estimated the feedback from the nine workers to cost about 20 Euros, but was concerned whether the workers had *"even evaluated the app or just thrown randomly points in the questionnaire."*

Many students from the bungaloo camp exhibited a naïveté towards what an acceptable pay for the feedback would be, and how to value crowd work in general. For example, P88 noted that *"the feedback helped streamline our app which I think is really valuable. A lot of the times small inconveniences can be a gamebreaker for products. Maybe 100€. I'm not quite shure [sic] how much feedback like that costs normally so its hard to put a concrete value to it."* P102 admitted to *"have no idea but I guess hundreds of euros, maybe closer to thousands."* P65 also noted that *"I don't have proper knowledge about how much such online workers charge but i think 20–30 euro would be fine."*

As is evident in some of the quotes above, students from the second camp often had a negative opinion of the feedback, but still considered paying a decent amount to the crowd workers. P96, for example, complained that *"all of the open ended questions were not useful,"* but thought the feedback was still worth 5€. In the same vein, P32 mentioned the feedback was *"really poor,"* but still considered paying *"10 EUR per worker."*

The primary contributor to this disaccord between perceived usefulness and monetary valuation was the student's mental image of the crowd worker as a trained professional with experience in usability and user experience (UX) testing. The suspected professionalism strongly affected the students' valuation of the crowdsourced feedback: *"I would only pay for one of the workers, maybe around 5-20 Euros depending on the expertise of the worker in this field"* (P9). The poor quality of the feedback did not dissuade the students from thinking that the workers deserved to be paid, often with extremely generous amounts: *"As they are professionals , if I consider the feedback on my application, the cost be between 100-200 Euros"* (P36). Similarly, P68 valued the feedback with *"150 [EUR] max,"* even though *"some [open-ended] feedback was only numbers like '3'."*

6 DISCUSSION

The majority of students in our study perceived peer feedback as more useful, more detailed, more specific, more effective, more actionable, and of greater fairness than crowdsourced feedback from MTurk. Further, the crowd workers spent less effort on writing their feedback, as evident in the high percentage of short answers. Peer feedback was more elaborate and contained less noise than the crowdsourced feedback.

In line with prior work [27], the qualitative analysis of the final questionnaire revealed that students perceived the affective tone (valence) of anonymous feedback to be slightly more positive compared to feedback from peers. Peer feedback was perceived to be much harsher and critical in tone. This, however, did not negatively impact the students' sense of fairness. On the contrary, students appreciated critical feedback, as long as it was elaborate, specific, well justified, and useful. We found, however, a gap in how student's perceived and valued the crowdsourced feedback which may contribute to a false sense of achievement in students, as discussed in the following section.

6.1 Perception Gap between the Monetary Value and Usefulness of the Feedback

Our data highlights that a student's perception of the monetary value of MTurk work may not be accurate. But since students likely have no experience with crowdsourcing, why is the students' perception of feedback value relevant and interesting? Why does it matter if they know how much to pay crowd workers?

From the students' individual perspective in the design task, the monetary value of feedback may indeed be irrelevant. From the perspectives of teaching, crowdsourcing research, and research funding, it is however exceedingly important to know that money

Table 4: Students' criteria for estimating the monetary value of crowdsourced feedback from MTurk.

Criteria and sub-criteria	Description
Perceived Usefulness	The usefulness of the feedback and its appropriateness for completing the design task.
• Perceived Quality	The overall quality of the feedback.
• Perceived Relevance	The relevance of the formative feedback for the students' design task.
• Perceived Effort	The effort the worker put into completing the task.
• Perceived Helpfulness	The amount of novel insights provided by the feedback to improve the design.
Professionalism	The assumed expertise and skill set of the worker for conducting usability and user experience tests.

on crowdsourcing is well spent as a monetary and pedagogical investment by the teaching organization. Knowing that students valued the feedback, and how they valued it, is therefore of particular importance to research and education institutions as well as teachers.

Over the years, many researchers have argued for fair work conditions on crowdsourcing platforms (e.g., [18, 39]). On the worker side, much of the discussion focuses on the imbalance in the dominant power structures on crowdsourcing platforms and on meeting a minimum wage for workers. From the requester's perspective, the crowdsourcing literature primarily focuses on the design of effective quality control mechanisms. Our work highlights a third side – that of the feedback receiver's qualitative experience of feedback.

The design space for crowdsourcing involving third parties has been only recently emerging. Examples include PledgeWork [20], a system for volunteers to donate their income from crowd work to a third party (a charity), revenue sharing (e.g., [8]), and the review of subcontracting microwork by Morris et al. [26]. In the classroom with three parties (i.e., the teacher, crowd workers and students), the requester of feedback (the teacher) may not be the receiver and beneficiary of the feedback (in this case, the students). This poses the question of how crowdsourced tasks for the collection of feedback should be priced. As the feedback mainly contributes value to the feedback receiver, the feedback receiver may be the right party to determine the feedback's value and hence its price. However, prior research on design feedback in the context of the classroom priced the crowdsourcing tasks without taking the feedback's value for students into consideration.

On the other hand, many students in our cohort were clearly unsure of how to value the crowdsourced feedback. Interestingly, these students still considered paying the workers a handsome amount of money for the feedback, even though the feedback was often of low usefulness. One reason for this gap in the perception and valuation of the feedback was that students conjured up an image of a professional and experienced online worker who is equipped with subject-specific expertise and an appropriate skill set. This conjured image, however, is in contrast to how MTurk is designed, as a marketplace for anonymous humans to complete short tasks irrespective of skills and expertise. Further, local salary standards obviously skew people's understanding about the monetary value of labor on MTurk.

All things considered, if teachers are to employ crowdsourced design feedback, teachers must educate their students about online work and its value. We contend that in the age of the gig economy, crowdsourcing and especially its valuation should be a fixed part of the Computer Science curriculum. Considering the strengths of employing crowdsourced feedback in the classroom, we concur with prior studies and argue that using crowdsourced feedback as a complement to traditional feedback mechanisms is useful for students.

6.2 Complementing Peers with the Crowd

Our work confirms the findings of prior studies that crowdsourced feedback is a good supplement to peer feedback [6, 37]. While the students in our study generally favored peer design feedback, they also discovered and acknowledged clear value in the crowdsourced design feedback – value that is impossible to be obtained in the classroom setting alone. For instance, the diversity of people providing the feedback was mentioned. In general, getting an outsider's perspective and feedback from diverse people from different cultures and backgrounds was valuable to the students. Crowdsourced design feedback was seen as a way to provide a reality check. Related to this, a few students stated that crowd workers did not *"hold back with their feedback"* (P67), whereas many perceived peer feedback as being *"somehow biased"* (e.g., P46, P100, and P105) even if they could not articulate this aspect in more detail. Further, the students reported that contrasting the feedback from the two different sources helped them identify the weaknesses and strengths in their designs, which allowed the students to set priorities and take appropriate action to address the weaknesses in their designs. More specifically, the students could distinguish between the two different groups of application users, and think how to improve the application for both groups. In the following section, we reflect on our own perspective as teachers of the HCI course.

6.3 Reflections from the Teachers' Perspective

The complementary value of crowdsourced feedback becomes a question of trade-offs between the added value to the learning versus the added burden to teachers and the monetary cost of the feedback. In our experience[1] there is certainly something of value in students seeing their designs being rated by other than familiar faces in the class: it is exhilarating and introduces an element of excitement to teaching. From the teachers' standpoint, utilizing MTurk as a feedback source was a refreshing experiment and provided us a chance to expose the students to feedback from people who are not their "friendly peers." Further, and while we have no data to back this up, we hypothesize that the knowledge of one's

[1]Several of the article's authors were involved in organizing this study's course.

design ending up online and being inspected by others than peers motivates the students to put in more effort.

Ultimately, teachers are the ones who choose to use crowdsourced feedback or not. In our case, for example, this will certainly not be the last time we employ crowd workers for additional feedback. Yet, several questions to consider remain. For instance, what is the correct way to determine the value of crowdsourced feedback? In the case of crowdsourcing platforms, the feedback is simply work that has a certain monetary value, dictated by the platforms' rules as well as the workers' and requesters' perceptions. The classroom as a context distorts this simple way of looking at value, however. In many institutions teaching does not have any extra budget available, and the question of value is not just about money. It is also a question of time and effort required to set it all up: collecting and distributing the feedback. We did this manually, and it certainly was not an insignificant amount of work. Not all teachers are equipped with the skills and knowledge required to crowdsource feedback from the crowds, and learning such skills could be difficult. There is a dearth of tools that would ease the setup for non-technically savvy teachers, and given how different each case most likely is, it is challenging to even envision sufficiently generic and easy-to-use tools for crowdsourcing feedback that would work across all types of classes.

Further, with teaching one has to always consider the quality of feedback. Instructors supposedly are experts in the subjects they teach, and the students can thus trust the feedback and its quality. With crowds, there is always an uncontrolled element of uncertainty involved. Should we trust that the crowd feedback, collected and distributed as-is, is of sufficient quality, or should the instructors vet and double-check everything? Would this, then, lead to loss of authenticity in the crowd feedback, even if authenticity is one of the perks of it? We believe that a fully transparent approach might work best here; Collecting the feedback from crowds but at the same time explaining the students exactly what they are in for and from what type of workforce.

We envision students to take a more active role in collaboratively working with crowds in the classroom, and we have already taken this approach in our own teaching. Our work may inspire the design of pedagogical tools and interventions (e.g., a set of rubrics), using the criteria that we outline in the paper as a framework. For example, we now encourage students to define their own set of rubrics and guiding questions for collecting feedback as part of their learning experience. To this end, we give recommendations for using crowdsourced feedback in the classroom in the next section.

6.4 Recommendations for Crowdsourcing Design Feedback in the Classroom

In this section, we consolidate recommendations that are common knowledge in the crowdsourcing community and available in the crowdsourcing literature (e.g., [6]), but may be novel for teachers in the HCI community who are less familiar with crowdsourcing and wish to use crowdsourcing as a source of feedback in the classroom.

6.4.1 Set the right expectations in students. We contend that teachers should educate their class about crowdsourcing (and crowd work in general), and especially about the working conditions on crowdsourcing platforms, and how such work is commonly priced. To effectively evaluate the crowdsourced feedback, students must be aware of the strengths and weaknesses of the crowdsourcing model. In the context of a course that provides crowdsourced feedback to students, educators should collect demographic data about the crowd workers' educational and professional background. This demographic data may help students to adjust their conflated expectations and arrive at a more realistic valuation of the crowdsourced feedback.

6.4.2 Set the right expectations in crowd workers. According to feedback intervention theory, if feedback signals that an effort falls short of an expected standard, learners become motivated to increase their efforts to attain the standard [21]. To implement this mechanism successfully, the standard and expectations must be clearly explained to the feedback provider. The crowd workers should be made aware that their feedback is given to students as formative, not summative feedback. Feedback requesters should define clear success criteria for the crowd workers and provide a definition of what defines good feedback. Structuring feedback with rubrics, scaffolds, guiding questions, and other structured work flows may help increase the quality of the feedback in this regard [2, 3, 44]. In particular, requesters of crowdsourced feedback should pay attention to how the criteria that affect the students' perception of quality and fairness shape the students' experience of feedback. Critical and harsh feedback, for instance, was not necessarily perceived as unfair by the students.

6.4.3 Apply crowdsourcing best practices of quality control. To effectively complement each other, both sources of feedback must contribute some valuable insights. To this end, typical qualification criteria as employed in our study (i.e., 95% past acceptance rate and 100 completed HITs) did not prove to be enough to motivate the MTurk workers to provide good feedback. To this end, the quality of the crowdsourced feedback could be elevated by filtering responses (post data collection). This is indeed a common practice with crowdsourced data collection. In our study, we found that the open-ended feedback provided by the crowd workers markedly improved when short responses were removed. The relevance of the feedback and its usefulness can be expected to improve if these responses are discarded.

6.4.4 Use an appropriate feedback modality and consider feedback aggregation. Another consideration is how feedback should be collected and provided to students. As is evident in a number of crowd feedback systems (e.g., [29, 41, 43]), visualization and aggregation of feedback supports the feedback receiver in making sense of the feedback. Research on mechanisms for aggregating crowdsourced design feedback is only recently emerging (e.g., [43]). Our study found that students in particular valued diversity in the responses and appreciated the direct contrast between the feedback from the two sources.

6.5 Limitations

We acknowledge limitations in our study. First, the feedback receiver's pedagogic literacy affects how feedback is evaluated [32]. Our findings may therefore be specific to the class and we do not claim that the findings generalize to other cohorts of students or other study subjects.

Second, the subjective experience of feedback is influenced by a number of factors. For instance, the order in which feedback is received may affect the perception of feedback [27, 40]. Nguyen et al. found that framing feedback for a writing task with positive affective language had a positive effect on work quality [27]. In Wu and Bailey's study, participants were more motivated and perceived the feedback as most favorable when negative feedback was given after positive feedback [40]. In our study, we did not control the order of feedback. Students explored the contents of the feedback package on their own terms. We argue that our study setup is not unrealistic, as it aligns well with the microtask crowdsourcing model found on MTurk and reflects how feedback could practically be provided in the classroom.

Third, a limitation of the investigation into the monetary valuation of crowdsourced design feedback is that while the questionnaire item specifically asked the students to estimate the value of the crowdsourced feedback as a whole, some students may still have estimated the value per worker. However, the students who elaborated on their answer typically mentioned whether their estimate was per worker or for the whole of the crowdsourced feedback, as discussed in Section 5.3.

Fourth, knowledge of the feedback source may have influenced the students' interpretation of the feedback. The choice to disclose the source was, however, imperative for teaching students the value of the two feedback sources. We believe revealing the source at the end of the study would have created tension among the students. Further, peer feedback was not assigned a value in our study. Peer feedback in our context (i.e., teaching an HCI class) is a mandatory part of the course and a traditional element of teaching and learning. In our institution's ethical stance, peer feedback can and should not be assigned a monetary value, and taxation law prohibits us from handing out money to students. We did, however, not see signs of insincerity in the students' answers (after removing the few unrealistic responses – see Section 5.3.1). We acknowledge that alternative study designs could have been explored.

Last, we did not control the students' familiarity with crowdsourcing. Students received a general introduction to crowdsourcing in the lectures (but without going into detail about the work reality on crowdsourcing platforms). A few students had prior knowledge of the work conditions on crowdsourcing platforms and were thus able to accurately estimate the price of crowdsourced work. The majority of the students, however, was new to crowdsourcing, as evident in their responses to our survey.

7 CONCLUSION

In this work, we provided a detailed empirical study of how students perceive and value crowdsourced feedback, as well as how crowdsourced feedback compares to peer feedback in the classroom. We found that students preferred peer feedback as it was perceived as more useful, fair, effective, and actionable. Additionally, our investigation of the students' monetary valuation of the crowdsourced feedback revealed quality, relevancy, effort, and helpfulness as important factors that shape the value of the feedback for students. We found clear evidence that some students were naïve toward the work conditions on crowdsourcing platforms and how such work is priced. The monetary valuation of the crowdsourced feedback by these students was strongly shaped by a mental image of the worker as a trained professional. Ultimately, we believe that crowdsourced design feedback in HCI teaching provides a great way to complement peer feedback, as long as student expectations are calibrated adequately.

ACKNOWLEDGMENTS

We thank the students of the HCI course and the workers from Amazon Mechanical Turk who contributed their feedback in this study. This research is partially enabled by the GenZ strategic profiling project at the University of Oulu, supported by the Academy of Finland (project number 318930).

REFERENCES

[1] Jennifer Carr. 2011. Providing Effective Feedback. , 15–17 pages. https://ssrn.com/abstract=1919096

[2] Amy Cook, Steven Dow, and Jessica Hammer. 2020. Designing Interactive Scaffolds to Encourage Reflection on Peer Feedback. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1143–1153. https://doi.org/10.1145/3357236.3395480

[3] Amy Cook, Jessica Hammer, Salma Elsayed-Ali, and Steven Dow. 2019. How Guiding Questions Facilitate Feedback Exchange in Project-Based Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300368

[4] Deanna P. Dannels and Kelly Norris Martin. 2008. Critiquing Critiques: A Genre Analysis of Feedback Across Novice to Expert Design Studios. *Journal of Business and Technical Communication* 22, 2 (2008), 135–159. https://doi.org/10.1177/1050651907311923

[5] Biplab Deka, Zifeng Huang, Chad Franzen, Jeffrey Nichols, Yang Li, and Ranjitha Kumar. 2017. ZIPT: Zero-Integration Performance Testing of Mobile App Designs. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. Association for Computing Machinery, New York, NY, USA, 727–736. https://doi.org/10.1145/3126594.3126647

[6] Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/2470654.2470686

[7] Josemaria Elizondo-Garcia, Christian Schunn, and Katherina Gallardo. 2019. Quality of Peer Feedback in Relation to Instructional Design: A Comparative Study in Energy and Sustainability MOOCs. *International Journal of Instruction* 12, 1 (2019), 1025–1040.

[8] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 132 (Oct. 2020), 24 pages. https://doi.org/10.1145/3415203

[9] Gerhard Fischer, Kumiyo Nakakoji, Jonathan Ostwald, Gerry Stahl, and Tamara Sumner. 1993. Embedding Computer-Based Critics in the Contexts of Design. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/169059.169133

[10] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[11] Tyler Hamby and Wyn Taylor. 2016. Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples. *Educational and Psychological Measurement* 76, 6 (2016), 912–932. https://doi.org/10.1177/0013164415627349

[12] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[13] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487

[14] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[15] Simo Hosio, Jorge Goncalves, Theodoros Anagnostopoulos, and Vassilis Kostakos. 2016. Leveraging Wisdom of the Crowd for Decision Support. In *Proceedings of*

the 30th International BCS Human Computer Interaction Conference: Fusion! (HCI '16). BCS Learning & Development Ltd., Swindon, UK, Article 38, 12 pages.

[16] Jeff Howe. 2006. The Rise of Crowdsourcing. *Wired Magazine* 14, 6 (2006), 1–4.

[17] Julie S. Hui, Amos Glenn, Rachel Jue, Elizabeth M Gerber, and Steven P. Dow. 2015. Using Anonymity and Communal Efforts to Improve Quality of Crowdsourced Feedback. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing*, Elizabeth Gerber and Panos Ipeirotis (Eds.). AAAI Press, Palo Alto, CA, USA, 72–82.

[18] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 611–620. https://doi.org/10.1145/2470654.2470742

[19] Hyeonsu B. Kang, Gabriel Amoako, Neil Sengupta, and Steven P. Dow. 2018. Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174180

[20] Keiko Katsuragawa, Qi Shu, and Edward Lank. 2019. PledgeWork: Online Volunteering through Crowdwork. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300541

[21] Avraham N. Kluger and Angelo S. DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119, 2 (1996), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

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

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

[24] Amna Liaqat, Cosmin Munteanu, and Carrie Demmans Epp. 2020. Collaborating with Mature English Language Learners to Combine Peer and Automated Feedback: a User-Centered Approach to Designing Writing Support. *International Journal of Artificial Intelligence in Education*. A Festschrift in Honour of Jim Greer. (2020), 1–42.

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

[26] Meredith Ringel Morris, Jeffrey P. Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. 2017. Subcontracting Microwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1867–1876. https://doi.org/10.1145/3025453.3025687

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

[28] Jonas Oppenlaender and Simo Hosio. 2019. Towards Eliciting Feedback for Artworks on Public Displays. In *Proceedings of the 2019 ACM Conference on Creativity & Cognition (C&C '19)*. ACM, New York, NY, USA, 562–569. https://doi.org/10.1145/3325480.3326583

[29] 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, Article 76 (6 2020), 28 pages. https://doi.org/10.1145/3394978

[30] Viktoria Pammer-Schindler, Erik Harpstead, Benjamin Xie, Betsy DiSalvo, Ahmed Kharrufa, Petr Slovak, Amy Ogan, Joseph Jay Williams, and Michael J. Lee. 2020. Learning and Education in HCI: A Reflection on the SIG at CHI 2019. *Interactions* 27, 5 (Sept. 2020), 6–7. https://doi.org/10.1145/3411290

[31] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (Dec 2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[32] Margaret Price, Karen Handley, Jill Millar, and Berry O'Donovan. 2010. Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education* 35, 3 (2010), 277–289. https://doi.org/10.1080/02602930903541007

[33] Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal* 24, 1 (2012), 92–96. https://doi.org/10.1080/10400419.2012.650092

[34] D. Royce Sadler. 1989. Formative Assessment and the Design of Instructional Systems. *Instructional Science* 18, 2 (Jun 1989), 119–144. https://doi.org/10.1007/BF00117714

[35] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (Jan 2017), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

[36] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1243–1252. https://doi.org/10.1145/1124772.1124960

[37] Helen Wauck, Yu-Chun (Grace) Yen, Wai-Tat Fu, Elizabeth Gerber, Steven P. Dow, and Brian P. Bailey. 2017. From in the Class or in the Wild? Peers Provide Better Design Feedback Than External Crowds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5580–5591. https://doi.org/10.1145/3025453.3025477

[38] Robert Philip Weber. 1990. *Basic Content Analysis* (2nd ed.). Sage, Newbury Park, CA, USA. https://doi.org/10.4135/9781412983488

[39] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP '19)*. AAAI, Palo Alto, CA, USA, 197–206.

[40] Y. Wayne Wu and Brian P. Bailey. 2017. Bitter Sweet or Sweet Bitter? How Valence Order and Source Identity Influence Feedback Acceptance. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. Association for Computing Machinery, New York, NY, USA, 137–147. https://doi.org/10.1145/3059454.3059458

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

[42] Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. 2015. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1637–1648. https://doi.org/10.1145/2675133.2675140

[43] Yu-Chun Grace Yen, Joy O. Kim, and Brian P. Bailey. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376380

[44] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1017. https://doi.org/10.1145/2818048.2819953