Novices Who Focused or Experts Who Didn't? How Effort and Expertise Cues Affect Judgments of Crowd Work

Y. Wayne Wu and Brian P. Bailey
Department of Computer Science
University of Illinois
Urbana, IL 61801
{yuwu4, bpbailey}@illinois.edu

ABSTRACT
Crowd feedback services offer a new method for acquiring feedback during design. A key problem is that the services only return the feedback without any cues about the people who provided it. In this paper, we investigate two cues of a feedback provider – the *effort* invested in a feedback task and *expertise* in the domain. First, we tested how positive and negative cues of a provider's effort and expertise affected perceived quality of the feedback. Results showed both cues affected perceived quality, but primarily when the cues were negative. The results also showed that effort cues affected perceived quality as much as expertise. In a second study, we explored the use of behavioral data for modeling effort for feedback tasks. For a binary classification, the models achieved up to 92% accuracy relative to human raters. This result validates the feasibility of implementing effort cues in crowd services. The contributions of this work will enable increased transparency in crowd feedback services, benefiting both designers and feedback providers.

Author Keywords
Crowdsourcing; design; feedback; creativity.

ACM Classification Keywords
H.5.3 [Information Interface and Presentation]: Group and Organization Interfaces – Collaborative computing.

INTRODUCTION
Crowd feedback services offer a new method for acquiring formative feedback during the iterative design process [37]. The services utilize online crowds as a simulated audience to collect, aggregate, and present their interpretation of a design [11, 21, 37]. Relative to soliciting feedback from peers and online communities, the benefits of these services include the ability to acquire feedback on-demand without burning social capital or needing online reputation [23], the integration of scaffolding to boost feedback quality [21], and access to a diverse and scalable audience [37]. Crowd feedback services can be used to acquire feedback on Web, product, and interaction designs, among other genres.

An empirical study of one representative service found that crowd feedback helps designers improve their designs in an iterative process [38]. However, in that study and other work [21], designers reported wanting to know more about the *providers* giving the feedback. This information could be used for assessing the credibility of responses, weighing conflicting viewpoints, and prioritizing suggestions. The problem is that existing crowd feedback services only show the feedback, without any information about the providers. One key reason is that there is little empirical knowledge about what information these services should display.

In this paper, we draw on social transparency theory [35] to study how presenting two critical cues about the providers – their effort and expertise – affect the perceived quality of their feedback. For this paper, *effort* is how much energy a provider invests in performing a crowdsourced task. For example, for a design feedback task, effort may include how long a provider views the design, length and number of revisions of the text, and the precision of the annotation. *Expertise* refers to the level of domain knowledge. While expertise has been studied for assessing online content [19, 26] and effort has been cited as critical for assessing crowd work [32], our work synthesizes and investigates these two cues for interpreting crowdsourced design feedback.

Our investigation of these cues consisted of two studies. In the first study, we generated an authentic dataset of design feedback and asked human raters (N=2700) to review each response and rate its perceived quality. In the rating interface, we manipulated a block of text giving positive and negative cues of the effort and expertise of the feedback provider. Results showed that both cues affect judgments of perceived quality relative to a baseline condition (up to 21% difference), but mainly for the negative manipulations. Surprisingly, we also found that indicating effort affects the perceived quality ratings as much as indicating expertise.

The results argue for implementing these cues in crowd feedback services, e.g., to help designers interpret and differentiate the feedback. For expertise, system designers

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858330

can choose between several existing methods (e.g. [28, 30, 33]). However, implementing cues of effort is challenging because it is a task-specific behavior and there has been little research aimed at measuring it for crowdsourced tasks.

Our second study therefore addressed this gap. We first collected behavioral traces of providers performing three feedback tasks. Through a software tool that we developed, human raters viewed replays of the workers performing the tasks and rated the perceived effort. A novel aspect of our replay tool is that it masked the characters during text entry to focus attention on the *behavior* rather than the content. Statistical models were built to learn mappings from the behavioral data to the perceived effort ratings. For a binary classification of effort, the models achieved 92% accuracy. This outcome demonstrates the feasibility of implementing effort cues within feedback services and other crowd tasks.

The contributions of this work are (i) empirical evidence showing how effort and expertise cues affect interpretations of crowdsourced design feedback; (ii) results indicating the feasibility of using statistical models for implementing effort cues in crowd work; and (iii) a general method and software tools for investigating effort in online work. Our contributions will enable increased social transparency in crowd feedback services, benefiting both the designers and the feedback providers.

RELATED WORK

We describe how our work is original relative to prior studies of assessing online content and situate it in context of social transparency theory. We also contrast our use of behavioral data for modeling effort in crowdsourced work.

Assessing Online Content

Relevant cues provided in the information environment can help users better judge credibility [7, 26], weigh conflicting views [19], make decisions [3], and prioritize suggestions. An important and generalizable cue is the expertise of the content's author and researchers have studied how this cue relates to content assessment [7, 19, 26]. For example, Liao and Fu found that online comments showing indicators of high expertise were selected by users for reading more often than comments without such indicators [19]. In a large-scale study, Fogg et al. found that the presence of expertise cues related to more favorable perceptions of website credibility [7]. In our work, we are also interested in how expertise cues affect the assessment of online content. However, our work tests the effects of expertise cues for evaluating online design feedback, a unique type of content; how cues of the provider's expertise interact with cues of his or her effort for the assessments; and how these effects are mediated by the intrinsic quality of the feedback.

Researchers have also identified the need to consider the effort of the content's author when assessing its quality, especially for crowdsourced work [32]. For example, one crowdsourcing study reported that up to 50% of the responses were of poor quality due in part to workers not investing sufficient effort into the task [16]. Our work is the first to study how explicit cues of effort affect the assessment of the quality of crowd work. Many other cues have also been studied for assessing content online [20, 22, 25, 27], but our focus is on studying the cues of effort and expertise in a crowdsourcing context.

Social Transparency in Online Work

Social transparency is defined as the availability of social meta-data surrounding information exchange [35]. Receiving design feedback is one form of information exchange and therefore it can be situated in the framework of social transparency. Though social transparency points to many attributes of social meta-data, our work considers two: expertise and effort. Expertise can be regarded as an attribute of identity transparency because it reflects a person's knowledge in the domain of interest. Effort can be regarded as an attribute of content transparency because it relates to the provider's behavior around the creation of the feedback. We prioritized these two attributes because expertise has been shown to be important for assessing online content [19] while effort has been described as being critical for interpreting crowdsourced work [32].

Prior work indicates that increasing social transparency can improve the quality of crowdsourced work [12] and affect impressions of those who performed the work [22]. However, the focus of these prior studies was to increase the transparency between crowd workers whereas we are increasing the transparency between a designer (requester) and the feedback providers (workers) to help the designer better interpret their responses. We are also using different transparency cues that are relevant to a design context.

Modeling Crowdsourced Work

There is growing interest in modeling the behavior of crowd workers for improved quality control [31, 32], task pricing [4], and activity history [22], among others. For instance, Rzeszotarski and Kittur have shown that behavioral traces of workers can be leveraged to predict response quality [32]. The authors further showed that models of behavior could be used to cluster workers who share similar patterns of work [31]. In contrast, part of our work tests how well models of behavior can be used to predict perceived *effort* rather than response quality. To determine a fair price for crowd work, Cheng and Bernstein leverage the objective performance data of workers to measure the intrinsic difficulty of a task and use it to set the task's price [4]. We are using the way workers perform a feedback task, which is subjective and open-ended, to model the perceived effort invested by the worker rather than the intrinsic difficulty of the task with the goal of helping designers better assess the feedback. Researchers have also developed models of crowdsourced work for completing work under budget or time constraints [9] or recommending tasks [1]. In contrast, our focus is on using behavioral traces to model perceived effort and to study its impact on judgments of the quality of crowd work.

STUDY 1: METHODOLOGY

Our first study examined how providing explicit cues of effort and expertise affects judgments of design feedback. The study addressed two fundamental research questions:

- RQ1: How do explicit cues of a provider's effort and expertise affect the perceived quality of the feedback provided for a design?
- RQ2: How are the effects of these cues mediated by the intrinsic quality of the feedback?

There is a large space of potential cues (social activity, demographics, geography, etc.), but we prioritized two to keep the study tractable. Expertise was included because it has been shown to be important for assessing online content [19] and because designers have reported wanting this cue for interpreting crowd feedback [37]. Effort was included because it has been previously described as important for assessing the quality of crowdsourced responses [32].

Experimental Design

To answer these questions, we conducted a full-factorial, between-subjects experiment. The factors were Effort (High vs. Low vs. Not Given) x Expertise (High vs. Low vs. Not Given) x Intrinsic Quality (low=1 to high=5), giving a 3x3x5 design. The experimental design and manipulations drew from similar studies testing how informational cues affect judgments in other domains (e.g. [2, 3, 26, 29]).

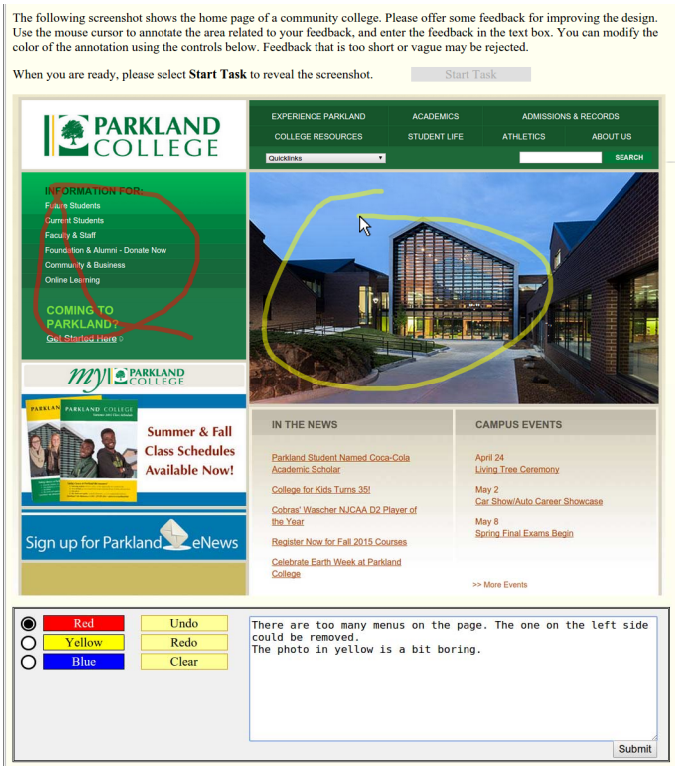


Figure 1. The user interface for collecting authentic design feedback from the crowd. Providers reviewed the Web design and left feedback using the edit box and annotation tools.

Design Feedback Dataset

For the experiment, we generated an authentic dataset of design feedback from feedback providers and developed an intrinsic quality score for each response. The design was the home page of a community college (http://parkland.edu). It was selected because its content should be familiar to a general audience, it was not too complex, and there were many opportunities for design improvements.

Feedback providers were recruited from Amazon.com's Mechanical Turk. A HIT was posted asking the providers (workers) to inspect the page and describe how it could be improved. The instructions also stated that feedback that was too short or vague would be rejected. As shown in Figure 1, the feedback collection interface included a screen capture of the page, an edit box for entering text, and a free-form ink tool (added via JavaScript) for annotating image regions corresponding to the text. The ink tool supported multiple colors and operations such as undo and clear. The

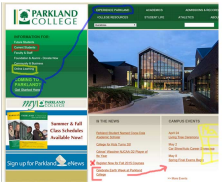
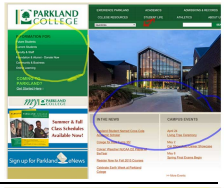
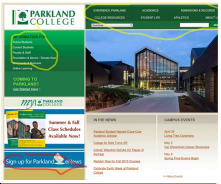
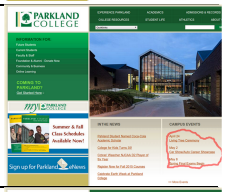
IQ	Feedback Text	Annotation
5 (High)	The "current students" section could easily be placed under the "student life" section. That's where I always found it for the homage of my university. I would also place "online learning" under academics … What is the difference between the two blue areas I circled? … Register Now should be at the very top of the page. …	
4	Red - I don't see the relation in the quick links and having other "quick" links at the top. Yellow - Font could be larger to be more appealing. Blue - "Important" information seem bland; could be presented a little more interactive to increase circulation.	
3	Make the information bar and the top bars more visible. They are boring and need to be more interesting to the student. Also get rid of the sign up for parkland enews since that should put on a page for current students. make your page more demanding for the prospective students.	
2	In ability to add campus events to an existing calendar, such as google calendar or iphone.	
1 (low)	Styling is not good.	

Figure 2. The authentic design feedback sampled at each level of intrinsic quality for Study 1. Some text in the top row (IQ 5) was omitted for brevity. Participants rated the perceived quality of the feedback with manipulations of the effort and expertise of the provider who supposedly left each response.

interface was designed to simulate existing crowd feedback services. Sixty pieces of feedback were collected, each from a different provider. A provider received $0.35 (US) and was required to have a 95% prior approval rating.

Three judges with experience in HCI were recruited from our institution to review the design and then rate the quality of each piece of feedback. The judges had no affiliation with this research project. The rating was performed on a 5 point Likert scale from 1 (lowest quality) to 5 (highest). For calibration, each judge first reviewed a sample of the feedback at different quality levels based on our own analysis and was encouraged to use the full range of the scale. A judge viewed the feedback online, one response at a time, and entered ratings in an online spreadsheet shown on a second monitor. A judge could review the feedback and modify the ratings until satisfied. Each judge completed the ratings in about one hour and received $15 (US).

Once the ratings were collected, we averaged the three ratings for each feedback response and rounded to give the final classification, or *intrinsic quality score*. On the scale of 1 to 5, the distribution of the classifications was 16, 14, 16, 12, and 2 respectively. Krippendorff's alpha, a measure of reliability for multiple raters and categories, was 0.71, which represents good agreement [14]. The feedback with higher intrinsic quality scores typically had more words (μ=633.0 for level 5 vs. μ=77.9 for level 1), suggested more improvements, and the suggestions were more specific and actionable. One feedback response from each level of intrinsic quality was randomly selected for the experiment. The feedback selected for the study is shown in Figure 2.

Experiment Interface

The rating interface showed a feedback response, a block of text about the feedback provider, and interaction for rating the perceived quality of the feedback. Perceived quality was rated on a scale from 1 (low) to 5 (high). See Figure 3.

The cues for effort and expertise were manipulated in the block of text, and followed a similar linguistic pattern. For expertise, the pattern was: "It is known that the person who left the feedback has $LEVEL knowledge of design." Effort followed a similar pattern: "It is known that the person who left the feedback invested $LEVEL effort to develop the feedback". $LEVEL was replaced with "minimal" and "significant" in the respective conditions. For example, if Low Effort was crossed with High Expertise, the block of text would read: "It is known that the person who left the feedback invested minimal effort to develop the feedback. It is also known that the person who left the feedback has significant knowledge of design." For a Not Given condition, the respective sentence was not included. The block of text was a manipulation in the experiment and was not related to the provider who actually left the feedback. The blocks of text for each level of effort and expertise were then replicated for the feedback representing the five levels of intrinsic quality. All 45 conditions were constructed a priori. When neither sentence was provided

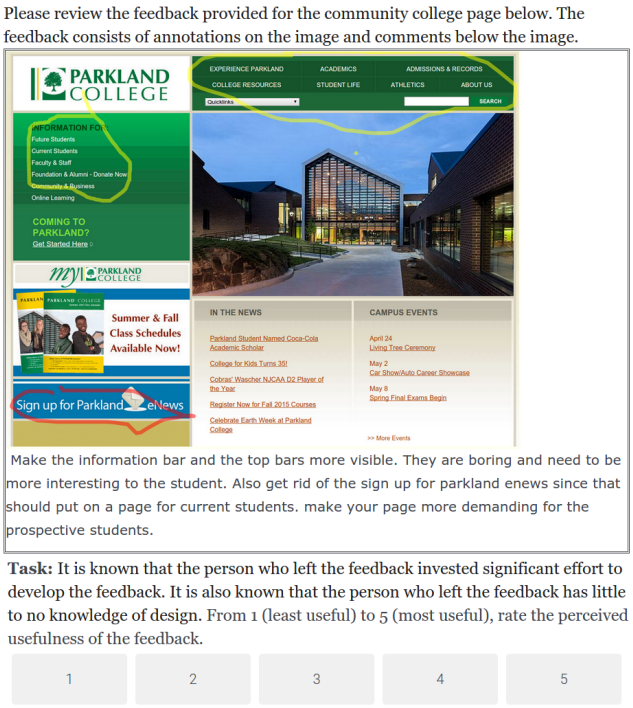


Figure 3. The interface for rating the perceived quality of the design feedback. It showed a feedback response, a block of text (manipulated) about the provider, and the rating interaction.

(i.e. effort not given and expertise not given), the participant only saw the feedback and the instructions for rating it, thereby serving as the baseline condition for the feedback at each level of intrinsic quality.

Participants

Participants (N=2700) were recruited from Mechanical Turk. Participants resided in the US (84.1%), India (12.1%), and 46 other countries (3.8%). We did not anticipate age or gender effects, and therefore did not collect this demographic data to minimize privacy concerns and to reduce the overall length of the task (HIT).

Procedure

Upon accepting the task, the participant was randomly assigned to one of the 45 conditions. Each condition was shown using the experiment interface previously described. The participant was instructed to review the feedback response for the design and rate its perceived quality (1 to 5). The manipulation of the text about the provider was integrated into the rating request, which was displayed below the design feedback and general task instructions. Participants therefore read the manipulation about the provider after viewing the feedback, but before rating it. Based on pilot testing, this placement achieved high likelihood that the manipulation was read. A participant received $0.10 (US) for performing the task. We piloted the task and used the mean task time to convert U.S. minimum wage to the payment. It was configured to require 95% prior approval and to allow workers to only participate

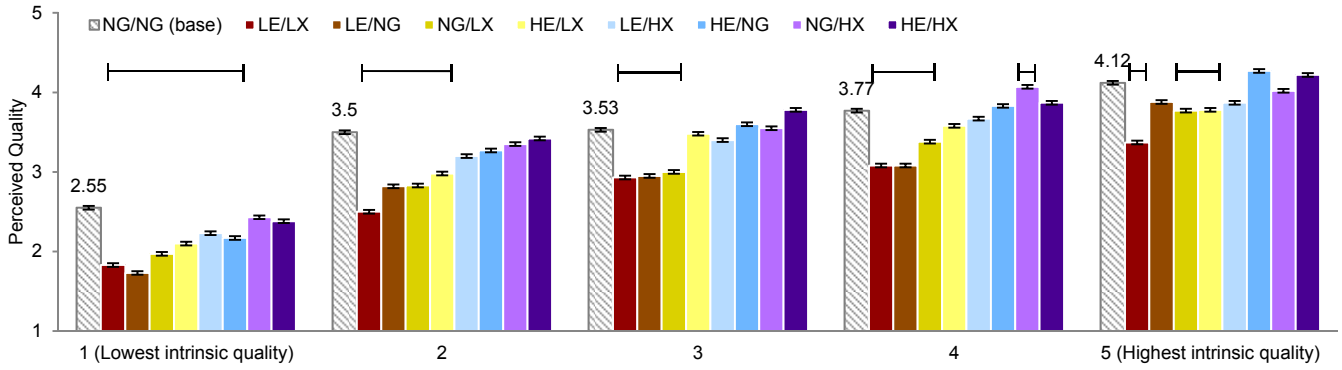


Figure 4. The graph shows the mean quality ratings across conditions (best in color, colors chosen from a color-blind safe palette). The x-axis clusters the conditions at the five levels of intrinsic quality. In each cluster, the left bar is the baseline and the bars are ordered from the most negative (left) to the most positive (right) cues given about the provider. For the legend, L=Low, H=High, E=Effort, X=Expertise, NG=Not given. For example, HE / LX is the High Effort / Low Expertise condition. Standard error = 0.023. The conditions under the horizontal markers are significantly different from the respective baseline ($p < .05$).

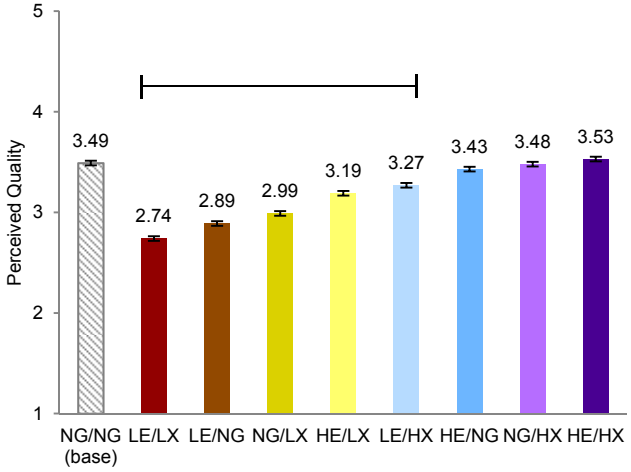


Figure 5. A graph of the perceived quality ratings collapsed across intrinsic quality. The legend is the same as figure 4.

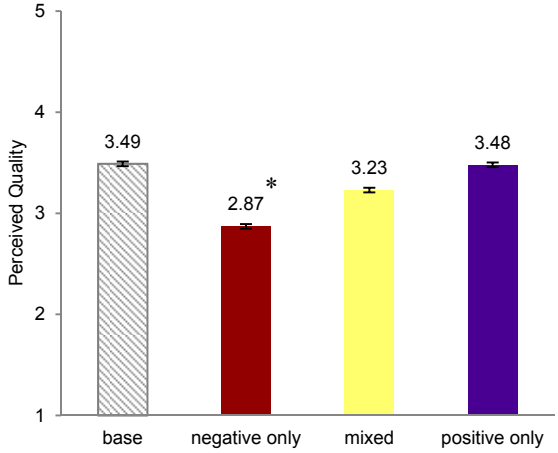


Figure 6. A graph of the perceived quality ratings comparing conditions showing only negative cues, a mix of positive and negative cues, and only positive cues about a provider. The 'negative only' group has a mean rating significantly lower than the other three groups shown ($p < .05$).

once. The batch was posted on AMT from June 29th to July 13th, 2015.

STUDY 1: RESULTS

We collected 3,081 ratings in total. 381 of the entries were discarded because they were incomplete, outliers (task time was two standard deviations away from the mean) or in excess of our target (60 per condition). The perceived quality ratings are shown in Figure 4 clustered by each level of intrinsic quality. Figure 5 summarizes the same ratings collapsed across intrinsic quality levels while Figure 6 groups the ratings by the valence of the cue conditions.

For example, in Figure 6, the 'negative only' group includes the Low Effort only, Low Expertise only, and Low Effort / Low Expertise conditions. Likewise, the 'positive only' group includes the High Effort only, High Expertise only, and High Effort / High Expertise conditions. The other cue conditions are included in the 'mixed' group.

Ratings were analyzed using a 3-way ANOVA with Effort, Expertise, and Intrinsic quality as factors. Bonferroni corrections were applied to pairwise comparisons to control for familywise error. The statistical results are summarized in Table 1. As a manipulation check, we analyzed the task time and found that raters spent more time on the task in the cue conditions (μ=57s) than in the control (μ=47s; $t(2698)=3.15$; $p<.01$). This confirms the cues were read.

Results showed an interaction effect between effort and expertise (row E:X in Table 1). Relative to the mean of the five baseline conditions (μ=3.49), signaling low effort regardless of expertise lowered the mean rating of perceived quality (μ=2.97, p<0.001). Signaling low expertise regardless of effort also lowered the mean rating (μ=2.97, p<0.001). However, when these cues were combined, the ratings of perceived quality were the lowest (μ=2.74, p<0.001). See Figures 5 and 6. In this case, the perceived quality of the feedback was reduced by 21% relative to the mean of the baselines. This pattern was consistent for the feedback at each level of intrinsic quality (see Figure 4) and is further supported by the lack of a three-way interaction.

Interestingly, the mean of the Low Effort only cue condition (μ=2.89) was lower than the Low Expertise only cue condition (μ=2.99, p<0.001). Knowing that a feedback provider did not invest effort into the task reduced perceptions of the work quality more than knowing that the provider only had minimal knowledge of the domain.

It was also surprising that when positive cues about a provider were shown (high effort or high expertise), the ratings of perceived quality were largely unaffected. For example, the mean rating in the High Effort / High Expertise condition (μ=3.53) was close to the mean of the baselines (μ=3.49, n.s.). Intrinsic quality also had a significant main effect on the ratings of perceived quality, as expected, but did not interact with the other two factors.

Discussion

The main results of the experiment were (i) showing cues of a provider's effort and expertise affect judgments of the quality of their feedback; (ii) negative cues had the largest effect on ratings of perceived quality, reducing ratings up to 21% relative to the baseline conditions, while positive cues had little impact; and (iii) effort cues were weighed similarly to expertise cues for judging feedback quality.

One immediate implication of our results is that a crowd feedback service should foreground the cues only when a provider exhibits unusually low effort or knowledge (the outliers) rather than when the cues are satisfactory (common case). This pattern would enable differentiation of the feedback responses without imposing the cognitive burden of having to interpret the cues for *every* response.

	df	SS	MS	F	p-value
Intrinsic quality	4	970.4	242.59	240.284	<0.0001**
Effort	2	89.3	44.64	44.214	<0.0001**
Expertise	2	96.7	48.37	47.909	<0.0001**
E:X	4	16.6	4.16	4.12	0.0025**
I:E	8	8.5	1.07	1.056	0.39
I:X	8	9.6	1.2	1.191	0.30
I:E:X	16	10.9	0.68	0.674	0.82
Residuals	2655	2680.5	1.01		

Table 1. Summary of three-way ANOVA applied to the perceived quality ratings. ** = significance at 0.01.

An interesting pattern in the results is that the negative cues led to reduced ratings of feedback quality without a commensurate increase in ratings for the positive cues. This pattern is consistent with the "negativity bias" in psychology showing that negative information influences evaluations more than positive information [14]. A potential mechanism driving this bias is that negative information is more memorable and seen as more non-normative [15]. The lack of influence of the positive cues may be due in part to participants assuming a certain level of effort and expertise on behalf of the providers in the baseline conditions. The positive cues were therefore only serving to reaffirm these assumptions.

Our results are consistent with a study by Carr and Walther reporting that positive cues did not lead to more favorable impressions of an evaluation target (job candidate) relative to a control [3]. However, the results of that work did not show a negativity bias which was present in our data. The difference may be due to different evaluation targets (job candidates in [3] vs. task outcomes in our study) or the style of cue presentation (inferred from anecdotes in [3] vs. explicitly given in our study). In [22], Marlow and Dabbish report that when evaluators are first shown positive cues of work history, their impressions of the worker are unaffected by subsequent work quality. In our study, the ratings were affected by the feedback (work) quality but not by the positive cues. The inconsistency could be due to differences in the experimental designs. For example, evaluators were shown an initial assessment of work history in [22] but not in our study. The targets were also different (people in [22] vs. task outcomes in ours). Other differences include the style, granularity, and number of cues presented, as well as different task domains. Future work is needed to tease apart these effects.

Affective priming theory potentially offers an alternative explanation of our results [18]. Signaling low effort or expertise could be considered a negative prime. We are skeptical of this explanation for two reasons. First, the cues about the provider (the prime) were read after reviewing the feedback, whereas priming typically requires seeing this information first. Second, if acting as an affective prime, one would expect to see increased ratings of quality for the positive cues, which were not present in the data collected. Additional aspects and implications of our study will be discussed in the General Discussion.

In our study, manipulating the effort and expertise cues was straightforward. But how could these cues be determined in a real-world crowdsourcing or other platform where the feedback exchange is typically remote and anonymous? For expertise, system designers could apply known techniques

such as performance-based assessments [33], aptitude tests [5, 28], or peer prediction [30]. For effort, however, there has been little research aimed at measuring it in a crowd context. Solutions such as self-reports may be ineffective due to strong biases against negative self-assessment (e.g. would workers really report that they made little effort on a task?), especially if linked to negative outcomes such as having the work rejected on a paid platform [4].

We therefore report on a second study which tests whether recording behavioral data collected during a design feedback task could be leveraged to predict the overall effort perceived by human raters. The study also contributes a new method for judging effort in an experimental setting.

STUDY TWO: METHODOLOGY

In study two, we explore the use of behavioral data for modeling perceived effort for design feedback tasks in a crowdsourcing context. The approach was to first collect behavioral data from providers leaving feedback for three designs and independent ratings of their perceived effort. We then built statistical models to learn mappings from features derived from the behavioral data to the ratings.

Behavioral Data Set

To collect behavioral data, we instrumented the feedback collection interface described in Study 1. See Figure 1. The interface was therefore used to collect both the feedback and the behavioral data for this study. Feedback and the associated behavioral data was collected for three Web designs; the home page of a community college (shown in Figure 1), an event organization site (http://evite.com), and a site for disseminating recorded talks (http://ted.com). The latter two sites along with some of the feedback provided are shown in Figure 7.

Using scripts added to the collection interface, we recorded the task behavior of the provider including mouse activity, keystrokes, interface and window actions, and start and end times for the main parts of the task. All events were time stamped. The scripts did not interfere with performing the task. Providers were not aware of this data collection.

The feedback collection interface was developed to aid the timings. For instance, after reading the general instructions, the provider had to select a button to reveal the design image and begin leaving feedback. This allowed us to record the preparation time (time from the onset of the task to the reveal of the image) and the design review time (from the reveal of the image to the first action). Sixty feedback responses were collected for each design, giving 180 total.

Replay Tool

Effort is how hard a provider works to give feedback on a design (e.g. how long did s/he view the design) and needs to be judged based on his or her behavior rather than on the content of the response. For instance, if a provider gave useful feedback but the content was blindly pasted from another source, then their effort on the task was minimal.

	Features	Explanation	Gain (rank)
Annotation	Bounding box area	Area of bounding box for the annotation	0.29 (7)
	Bounding box percent	Size of bounding box relative to the design image	0.28 (8)
	Stroke colors	Number of colors used for the annotation	0.18 (9)
	Covered area	Percent of pixels in bounding box covered by the annotation	0.18 (10)
	Max speed	Maximum cursor speed	0.15 (12)
	Average speed	Average cursor speed	0.14 (14)
	Max acceleration	Maximum cursor acceleration	0.13 (15)
	Strokes	Number of strokes used for the annotation	0.12 (17)
	Average acceleration	Average cursor acceleration	0.12 (18)
	Overpaints	Pixels painted by multiple strokes	0.06 (20)
	Undos	Number of stroke undos	0
Text Entry	Words	Word count of the text	0.64 (1)
	Characters	Character count of the text	0.59 (2)
	Deletions	Number of deletions during text entry	0.37 (4)
	Pauses	Number of pauses longer than two seconds when entering text	0.36 (5)
	Text ratio	Ratio of total annotation time to total time entering text	0.34 (6)
	Longest word	Length of the longest word in the text	0.25 (9)
	Average word length	Average word length of the text	0.17 (11)
	Insertions	Number of char insertions	0.14 (13)
	Typing speed	Average speed for typing the text	0.13 (16)
	Control actions	Total number of actions on the annotation control panel	0.07 (19)
Timings	Task time	Total time spent on the task	0.40 (3)
	Prepare time	Time taken to read the general description (from start of task to selection of “start feedback”)	0
	Image review	Time taken to review the design (from selection of “start feedback” until first action).	0
	Task review	Time from last action until the task is submitted.	0

Table 2. The features used for modeling perceived effort. The right column shows the information gain scores and (rank) for each feature for the binary classification. Within each category, the features are ordered by their rank.

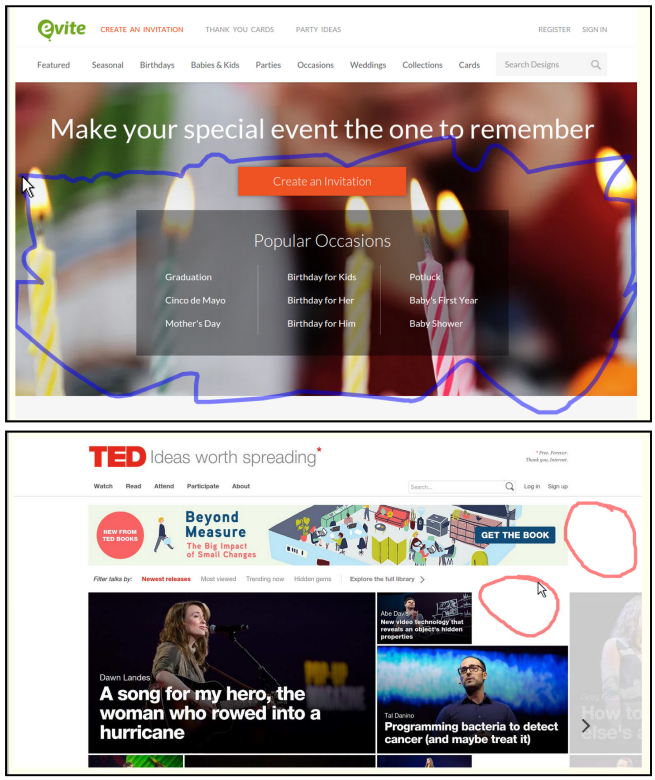


Figure 7. The two additional Web designs for which crowd feedback and behavioral data was collected for modeling perceived effort. The task instructions and annotation panel were omitted here but were the same as shown in Figure 1.

To enable effective judgments of effort, we built a tool that read the behavioral data and replayed (in the form of a video) the provider performing the task. To minimize influence from the feedback content, each character entered in the edit box was replaced with an 'x'. The purpose was to focus the judges' attention on the behavior (e.g. typing speed and content revision), rather than the content itself. If the idle time between actions was longer than five seconds, the tool enabled a "skip" button for jumping to the next recorded action. If the tool found a top-level window focus event during the idle time, it displayed a message that the provider switched to another window.

Judges

Three graduate students from our institution were recruited to judge the effort made by each feedback provider using the replay tool. The students were not affiliated with the project and did not participate in Study 1. Performing the ratings took about three hours and each judge was paid $85.

Procedure

After informed consent, judges received an overview of the study along with a description of *effort* - how much energy the provider invested in providing the feedback. Judges were presented with a sample of the replays to calibrate their ratings. The researchers informed the judges that they were free to develop their own criteria for judging the effort

Actual Rating \ Predicted Rating	1	2	3	4	5
1	**21**	7	0	1	0
2	3	**30**	10	0	0
3	0	11	**20**	4	1
4	0	4	10	**18**	7
5	0	0	0	7	**26**

	1	2
1	**91**	6
2	8	**75**

Table 3. The confusion matrices for predicting five levels and two levels of effort (1 = lowest effort).

observed but suggested considering aspects of the entered text, annotation, and duration. The judges rated the effort on a scale from 1 (low effort) to 5 (high effort). Ratings could only be made at the end of a replay and were entered into a spreadsheet shown on a second monitor. Judges were allowed to revisit replays and modify their ratings until satisfied. Each judge rated the effort of the 180 providers who gave feedback. The three ratings of each provider were then averaged to produce the final rating. Though the size of the data set was modest, it was sufficient for testing the feasibility of mapping the behavioral data to the ratings.

Features

We created 25 features from the behavioral data and these are shown in Table 2. The features were derived from discussions with the judges about what observed behaviors affected their ratings, our experience piloting the tasks and data collection, and prior work [8, 32]. The features are not exhaustive, but do provide a reasonable starting point for exploring statistical models of perceived effort. A feature vector was created for each feedback response.

STUDY 2: RESULTS

The rating distribution of the judges is shown in Figure 8 and was nearly uniform (μ=3.0). This validates that the feedback providers (workers) performed the feedback tasks with varying levels of effort. Krippendorff's alpha was 0.79, indicating good agreement among the judges. The fact that judges could agree on the effort observed suggests that statistical models could also learn the mappings.

All models were built using support vector machines in Weka 3.6 [18] and tested using ten-fold cross validation. Alternative statistical models including logistic regression, naive Bayes, and decision trees were also explored. These models produced similar results to what is reported below.

As a first step, we created models that learned mappings from the features to the five levels of effort. The results are summarized as a confusion matrix in Table 3 (left). The overall accuracy was 65%, precision was 0.65, recall was 0.64, and the F-measure was 0.64. From the table, the most egregious errors (e.g. actual low effort predicted as high effort or vice versa.) were rare. The accuracy was modest, but may be improved by training on a larger data set and extracting additional features from the behavioral data.

As an alternative to predicting the five levels of effort, we simplified the problem to a binary classification; *effortful* (ratings of 3, 4 or 5) and *not effortful* (ratings of 1 or 2). A binary classification would be easier to interpret and would be consistent with the two levels of effort manipulated in Study 1. A model was trained using the same data, but now for the binary classification. The accuracy was markedly improved (92%). Precision, recall, and the F-measure were all 0.92 and the results are shown in Table 3 (right). To determine which of the features contributed most to the classification, we performed feature selection using the information gain metric in Weka. The gain scores for each feature and their rank are shown in the right column of Table 2. The length of the text, total time on task, revisions made to the text, typing pauses to (presumably) review the design, and multiple colors in the annotation were among the features that contributed most to the classification.

GENERAL DISCUSSION

The results from Study 2 showed that it is feasible to model perceived effort for crowdsourced design feedback with up to 92% accuracy. There are at least two ways that crowd feedback services could apply this finding. One way is for the service to model the perceived effort of the providers and display the classifications (cues) for the designer. As discussed from the results in Study 1, the services should only foreground the cues when they would be most beneficial for differentiating the feedback (e.g. to signal unusually low effort or domain knowledge). A second way would be for the service to use the models to automatically reject low effort work and acquire more effortful responses. This approach would trade feedback generation time for a set of responses that are likely to be of higher quality. In fact, there was a strong positive correlation (Pearson's r = 0.82, $p<0.001$) between the ratings of perceived effort and quality for the feedback data set shared between the two studies in this paper.

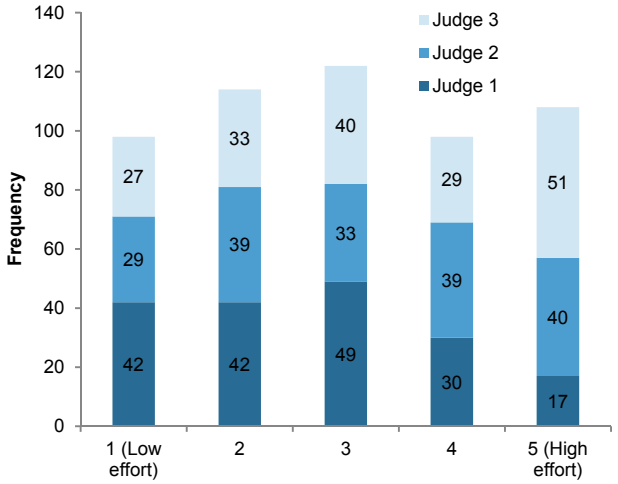


Figure 8. The distribution of ratings of the perceived effort invested by providers for the feedback tasks. There was good agreement overall and the distribution was close to uniform.

We modeled perceived effort for crowdsourced design feedback, but the approach generalizes to other tasks where users must judge subjective responses from the crowd. Such judgments can be found in work on crowd-based ideation [34], content summarization [17], and social data analysis [36]. Leveraging our methodology from Study 2, including the behavioral data collection and replay tools, researchers and system designers can build statistical models for their own tasks. It may also be possible to build a more general model by considering only lower-level features independent of the task type and training it on a larger data set [6]. A generalized model should also consider a worker's effort invested outside of the task environment, such as reviewing content on external sites to prepare for the task. One practical way to capture this effort is to prompt the worker (e.g. paste the URLs of sites reviewed) in the task interface. The model could then consider the content and estimated time spent on the external sites.

Logging behavioral data to make effort visible in a crowd service could raise privacy concerns. However, it could also lead to practices favorable for workers. For example, crowd services could enable users to pay bonuses based on the effort invested by workers. Workers may also improve their performance if they are able to view and reflect on their own effort, and possibly command higher pay with a reputation of effortful responses. The reputation could reflect the effort modeled for the individual tasks as well as the worker's propensity to perform sequences of tasks beyond the minimum. In addition to showing cues that summarize effort, a service might also show how the worker's logged behavior compares to other workers for the same task. This could be used to explain the cues or to improve performance by showing replays of effortful work as exemplars. Requesters can also benefit from making effort visible beyond aiding the interpretation of responses. For example, a large fraction of low effort responses may signal ineffective task design rather than malicious behavior.

The results from Study 1 showed that expertise cues also factor into judgments of feedback quality. Expertise measures could be implemented as qualification or screening tasks that gauge relevant aptitudes [5, 28], measure peer prediction ability [30] or apply performance-based assessments [33]. Future work could also explore extending models of user interface skill (e.g. [10, 13]) to model the domain expertise of crowd workers.

The manipulations of effort and expertise cues in Study 1 were achieved using specific phrasings of text. Researchers have already shown that different representations can have different influences over the evaluation of work quality [24]. This thread of research could therefore be extended to study different phrasings and granularities of the cues used in our study and in context of design feedback. It would be also interesting to study whether the cues always need to be displayed or only when needed to differentiate the feedback.

FUTURE WORK

Beyond the opportunities already discussed, we see several exciting directions for future work. One direction is to implement effort and expertise cues into an existing crowd feedback service and study how designers learn to leverage the cues for interpreting the feedback and making decisions, and how the benefits may be offset by the cognitive load or time needed for assessing the cues. It would also be interesting to study how the feedback providers react to the implementation of the cues. A second direction is to conduct experiments testing how a broader range of cues (beyond effort and expertise) affect judgments of subjective responses from the crowd. Ideally, researchers could tease out the optimal set of cues that provide the most benefit with the least computational and cognitive overhead. Another interesting direction is to compare perceived and self-reported effort or other similar cues. Finally, in contrast to showing cues of the feedback provider, it would be interesting to study how providing information about the designer may influence the quality of the feedback responses and the effort invested by the crowd.

CONCLUSION

Designers can leverage crowd services to quickly generate large quantities of feedback on in-progress designs. A key problem is that these services only present the feedback and do not display additional cues for helping designers weigh the responses. Our work has made three key contributions addressing this problem. First, we report empirical results showing that two cues – effort and expertise – affect the perceived quality of crowdsourced design feedback and that the effects are most pronounced for negative indicators of effort or expertise. Surprisingly, we also found that cues of effort affected the perceived quality as much as expertise.

Second, to fill a gap in the literature, we demonstrated the feasibility of building statistical models that use behavioral data to classify levels of perceived effort on feedback tasks. For binary classification the models achieved 92% accuracy relative to human raters. This modeling approach can be used to implement effort cues in existing feedback services. Finally, our work contributes a general method and tools for studying effort in online work. Others can also leverage this contribution to build and test statistical models of effort for other crowdsourced tasks. The tools can be downloaded from https://github.com/uiuc-crowd-research/chi2016. We are enthusiastic that our contributions can bring increased transparency to crowd feedback services, which will be beneficial for both the users of these services and those who perform the work.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under award NSF CMMI 14-62693.

REFERENCES

1. Vamshi Ambati, Stephan Vogel and Jaime Carbonell. Towards Task Recommendation in Micro-Task Markets. *Proceedings of the AAAI Workshop on Human Computation (HCOMP)*, 2011.
2. Michael H. Birnbaum and Steven E. Stegner. Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, *37* (1): 48-74. http://dx.doi.org/10.1037/0022-3514.37.1.48
3. Caleb T. Carr and Joseph B. Walther. Increasing attributional certainty via social media: Learning about others one bit at a time. *Journal of Computer-Mediated Communication*, *19* (4): 922-937. http://dx.doi.org/10.1111/jcc4.12072
4. Justin Cheng, Jaime Teevan and Michael S. Bernstein. Measuring Crowdsourcing Effort with Error-Time Curves. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2015, 1365-1374. http://doi.acm.org/10.1145/2702123.2702145
5. Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz and Scott Klemmer. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2011, 2807-2816. http://doi.acm.org/10.1145/1978942.1979359
6. James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee and Jie Yang. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction*, *12* (1): 119-146. http://doi.acm.org/10.1145/1057237.1057243
7. B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani and Marissa Treinen. What makes Web sites credible?: A report on a large quantitative study. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2001, 61-68. http://doi.acm.org/10.1145/365024.365037
8. Krzysztof Z. Gajos, Katharina Reinecke and Charles Herrmann. Accurate Measurements of Pointing Performance from In Situ Observations. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2012, 3157-3166. http://doi.acm.org/10.1145/2207676.2208733
9. Yihan Gao and Aditya Parameswaran. Finish them!: pricing algorithms for human computation. *Proceedings VLDB Endowment*, *7* (14): 1965-1976. http://dx.doi.org/10.14778/2733085.2733101
10. Arin Ghazarian and S. Majid Noorhosseini. Automatic detection of users' skill levels using high-frequency user interface events. *User Modeling and User-*

Adapted Interaction, *20* (2): 109-146. http://dx.doi.org/10.1007/s11257-010-9073-5

11. Michael D. Greenberg, Matthew W. Easterday and Elizabeth M. Gerber. Critiki : A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. *Proceedings of the ACM Conference on Creativity & Cognition*, 2015, 235-244. http://doi.acm.org/10.1145/2757226.2757249
12. Shih-wen Huang and Wai-tat Fu. Don't Hide in the Crowd! Increasing Social Transparency Between Peer Workers Improves Crowdsourcing Outcomes. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2013, 621-630. http://doi.acm.org/10.1145/2470654.2470743
13. Amy Hurst, Scott E. Hudson and Jennifer Mankoff. Dynamic detection of novice vs. skilled use without a task model. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2007, 271-280. http://doi.acm.org/10.1145/1240624.1240669
14. David E. Kanouse and L. Hanson. Negativity in evaluations. in *Attribution: Perceiving the Causes of Behavior*, General Learning Press, Morristown, NJ, 1972, 47-62.
15. David E. Kanouse. Explaining negativity biases in evaluation and choice behavior: Theory and research. *Advances in Consumer Research*, *11* (1): 703-708.
16. Aniket Kittur, Ed H. Chi and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2008, 453-456. http://doi.acm.org/10.1145/1357054.1357127
17. Aniket Kittur, Boris Smus, Susheel Khamkar and Robert E. Kraut. CrowdForge: crowdsourcing complex work. *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2011, 43-52. http://doi.acm.org/10.1145/2047196.2047202
18. Karl Christoph Klauer. Affective Priming. *European Review of Social Psychology*, *8* (1): 67-103. http://dx.doi.org/10.1080/14792779643000083
19. Q. Vera Liao and Wai-tat Fu. Expert Voices in Echo Chambers: Effects of Source Expertise Indicators on Exposure to Diverse Opinions. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2014, 2745-2754. http://dx.doi.org/10.1145/2556288.2557240
20. Kurt Luther, Kelly Caine, Kevin Ziegler and Amy Bruckman. Why It Works (When It Works): Success Factors in Online Creative Collaboration. *Proceedings of the ACM Conference on Supporting Group Work*, 2010, 1-10. http://doi.acm.org/10.1145/1880071.1880073
21. Kurt Luther, Jari-lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala and Steven P. Dow. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2015, 473-485. http://dx.doi.org/10.1145/2675133.2675283
22. Jennifer Marlow and Laura Dabbish. The Effects of Visualizing Activity History on Attitudes and Behaviors in a Peer Production Context. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2015, 757-764. http://dx.doi.org/10.1145/2675133.2675250
23. Jennifer Marlow and Laura Dabbish. From rookie to all-star: professional development in a graphic design social networking site. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2014, 922-933. http://doi.acm.org/10.1145/2531602.2531651
24. Jennifer Marlow, Laura Dabbish and Jodi Forlizzi. Exploring the Role of Activity Trace Design on Evaluations of Online Worker Quality. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2015, 1617-1620. http://dx.doi.org/10.1145/2702123.2702195
25. Jennifer Marlow, Laura Dabbish and Jim Herbsleb. Impression Formation in Online Peer Production: Activity Traces and Personal Profiles in GitHub. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2013, 117-128. http://dx.doi.org/10.1145/2441776.2441792
26. Elliott McGinnies and Charles D. Ward. Better Liked than Right: Trustworthiness and Expertise as Factors in Credibility. *Personality and Social Psychology Bulletin*, *6* (3): 467-472. http://dx.doi.org/10.1177/014616728063023
27. Miriam J. Metzger, Andrew J. Flanagin and Ryan B. Medders. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*, 413-439. http://dx.doi.org/10.1111/j.1460- 2466.2010.01488.x
28. Tanushree Mitra, C. J. Hutto and E. Gilbert. Comparing person- and process-centric strategies for obtaining quality data on Amazon Mechanical Turk. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2015, 1345-1354. http://dx.doi.org/10.1145/2702123.2702553
29. G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, *1*, 33-60. http://dx.doi.org/10.1080/14792779108401856

30. Drazen Prelec. A Bayesian truth serum for subjective data. *Science*, *306*, 462-466. http://dx.doi.org/10.1126/science.1102081

31. Jeffrey M. Rzeszotarski and Aniket Kittur. CrowdScape: Interactively Visualizing User Behavior and Output. *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2012, 55-62. http://dx.doi.org/10.1145/2380116.2380125

32. Jeffrey M. Rzeszotarski and Aniket Kittur. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2011, 13-22. http://doi.acm.org/10.1145/2047196.2047199

33. James Shanteau, David J. Weiss, Rickey P. Thomas and Julia C. Pounds. Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, *136* (2): 253-263. http://dx.doi.org/10.1016/S0377- 2217(01)00113-8

34. Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos and Steven P. Dow. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2015, 937-945. http://doi.acm.org/10.1145/2675133.2675239

35. H. Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird and Ruogu Kang. Social Transparency in Networked Information Exchange: A Framework and Research Question. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2012, 451-460. http://doi.acm.org/10.1145/2145204.2145275

36. Wesley Willett, Jeffrey Heer and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2012, 227-236. http://doi.acm.org/10.1145/2207676.2207709

37. Anbang Xu, Shih-wen Huang and Brian P. Bailey. Voyant : Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2014, 1433-1444. http://dx.doi.org/10.1145/2531602.2531604

38. Anbang Xu, Huaming Rao, Steven P. Dow and Brian P. Bailey. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2015, 1637-1648. http://dx.doi.org/10.1145/2675133.2675140