Bitter Sweet or Sweet Bitter? How Valence Order and Source Identity Influence Feedback Acceptance

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

ABSTRACT
Content creators are less receptive to feedback with negative valence, and such feedback is frequently received online. To address this problem, we propose a novel method that orders a set of feedback based on its valence; using the feedback with positive valence to mitigate the effects of the negative valence feedback. To test the method, participants (N=270) wrote a story for children based on a given illustration and then revised their story after receiving a set of feedback. The feedback set was delivered with different valence orders and with different source identity cues. We measured participants' affective states, perceptions of the feedback and its source, revision extents, and story quality. Our main result is that presenting negative feedback last improved content creators' affective states and their perception of the feedback set relative to placing the negative feedback in other positions. This pattern was consistent across all feedback source conditions. The work contributes a simple and novel way to order a set of feedback that improves feedback receptivity.

Author Keywords
Crowdsourcing; design; feedback; creativity support tools.

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

INTRODUCTION

Content creators can increasingly leverage crowdsourcing and social media to collect feedback during the iterative design process [42]. Online crowds are attractive because they can be used to quickly generate a large quantity of diverse feedback. A shared challenge is the presence of negative feedback. Though negative feedback can occur in face-to-face settings, providers are more likely to give harsh criticism and use language with negative valence in online environments [35]. Prior work shows that negative feedback accounts for approximately 20% of all messages exchanged on a popular community site [46]. Negative feedback can discourage content creators from collecting feedback online [42], and cause creators to form an unfavorable impression of the platform and deter future use. Negative feedback also impairs content creators' affective states, which in turn can inhibit creative thinking [1].

In formal learning environments, instructors recommend the use of mitigating language, such as praise or affirmation, before or after negative feedback to improve its receptivity [2,4,16,23,37]. Similar techniques are less applicable in online environments, where the feedback is often composed by multiple independent providers and platform designers have limited control over the composition process. Prior work has tested various methods that improve the positive valence of feedback, including the use of rubrics [44] and positive examples [11]. While these methods may help, they do not eliminate occurrences of negative feedback. One common solution among platform designers is to remove the negative feedback [12,33]. Despite its simplicity, this approach has several disadvantages. First, negative feedback may still contain constructive advice useful for learning and content improvement [8]. Second, removing feedback without consent may discourage the provider from participating further [18]. Third, feedback removal may be inapplicable in certain situations, such as when content creators have paid in advance for each piece of feedback.

In this paper, we test a novel approach that manipulates the order in which a content creator consumes multiple independent pieces of feedback based on their valence. Specifically, our work examined whether positive feedback could be used to mitigate the influence of the negative feedback in the set. If effective, the technique could be automated in existing feedback platforms using sentiment analysis [48]. Prior work also shows that cues about the source (provider) of the feedback can affect perceptions of the feedback [38]. Our experiment additionally examined how the perceived source of the feedback (peers vs. experts vs. anonymous) [7,24] affects content creators' reactions to the feedback, and how it mediates the effect of valence order.

We conducted an online experiment in which participants (n=270) wrote a children's story based on an illustration. Two days later they revised their stories based on a set of given feedback. The feedback set included two pieces of

C&C '17, June 27-30, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4403-6/17/06…$15.00
http://dx.doi.org/10.1145/3059454.3059458

feedback with positive valence and one piece with negative valence. The feedback was presented with different valence orders (negative first, negative last, and negative between) and source identity cues based on the experiment condition. We measured participants' affective states, perceptions of the feedback and its source, revision extent, and story quality.

Our results showed the later in the set content creators consume the negative feedback, the more effort they invest in the revision task and the more favorably they rate the set of feedback. Feedback source had no statistical effect on the measures. We also observed a strong gender effect. Female participants were more likely to accept feedback (reported applying it in the revised story) and the negative feedback causes a larger reduction in their affective states. Our work contributes deeper empirical understanding of how valence order and source identity can be used to improve feedback receptivity in online environments.

RELATED WORK

In this section, we discuss how prior work addresses negative feedback from online crowds. Also, we examine previous findings about how feedback valence and its source influence content creators' reactions to the feedback.

Negative Feedback in Crowd Feedback Frameworks

Researchers have explored various methods to address the issue of negative feedback in crowd platforms. One approach focuses on increasing the quality of each piece of feedback. Greenberg et al. use a high-quality feedback example to scaffold the writing process [11]. The positive examples encourage feedback providers to adopt similar writing style and avoid being too critical. However, the example provided may cause cognitive fixation during feedback composition, limiting the learning benefit of the feedback [17]. Prior work has also explored how pricing tasks in a paid work platform influences the quality of online peer production [20]. The results show that higher payments increase the speed and quantity of task outcomes but not the quality.

Aggregation is another approach for addressing negative feedback. Yatani et al. generate a word cloud from review content that surfaces the most popular phrases [41]. The aggregation gives content creators an overview of the feedback before they read the text. Other researchers improve quality by structuring the feedback generation process [19,40]. Feedback providers respond to specific prompts about a design, and the responses are summarized through visualizations. The summary creates a buffer between the content creator and the individual feedback. Though the scaffolding process may increase quality, it may also reduce the scope of the feedback. Finally, researchers have sought to place reflection activities within the feedback loop to improve feedback interpretation [43].

In comparison to these prior approaches, feedback ordering is a lightweight mechanism that can be incorporated into existing platforms without significant modification. It can also co-exist with other quality improvement strategies, presentation techniques, and reflection activities to further improve feedback receptivity.

Effects of Valence in Creative Work

Researchers have explored various factors that influence feedback receptivity, including granularity, modality, and timing [9,15,21]. For example, Sadler argues effective feedback needs to address the discrepancy between the current performance level and the desired goal [32]. Valence is another factor with a strong influence over content creators' reactions to the feedback received. Negative language can threaten one's ego and reduce feedback effectiveness [5]. Positive language typically improves feedback reception, task performance, and content creators' affective states [24]. Zhu et al. show negative feedback discourages participation in online production communities while positive feedback encourages participation [39,47].

Feedback valence is particularly important for creative tasks. Positive affective state, which can be affected by feedback, relates to improved creativity [1,30,45]. Researchers have explored workflows that utilize the effects of positive valence in creative work. Nguyen et al. modify the valence level of the feedback by inserting positive affective language at the beginning of the text [24]. Such a positive language "wrapper" was shown to improve feedback reception and writing quality. De Rooij & Jones show that displaying positive feedback in real time can encourage participants to generate more original ideas [29]. In comparison with these workflows, our approach of feedback reordering requires neither modification of the feedback content nor real-time generation that may be challenging on many platforms.

Prior work has explored how manipulating the valence of phrases within a single piece of feedback influences content creators. One prior study shows that delivering negative feedback with positive feedback framing increases its perceived usefulness and participants' confidence but does not affect participants' performance on a repetitive physical task [25]. Prior work has also shown that mitigating language increases participants' receptivity to feedback and their affective state [2,16]. In contrast, our work explores the effects of valence ordering in the context of multiple pieces of feedback and in the context of a creative writing task.

Source Identity Influence on Feedback Receptivity

The identity of feedback providers is a common factor that differentiates feedback platforms. Online work marketplaces, such as UpWork, allow content creators to identify and collect feedback from paid domain experts. Feedback from experts leads to greater improvement in technical skills than self-assessment [27]. Prior work also finds content creators are more likely to accept feedback from experts because of their high perceived credibility [7].

On the other hand, collecting feedback from peers is also gaining popularity among content creators [13]. Although less experienced than experts, peers are typically more accessible for feedback collection. Prior work also shows

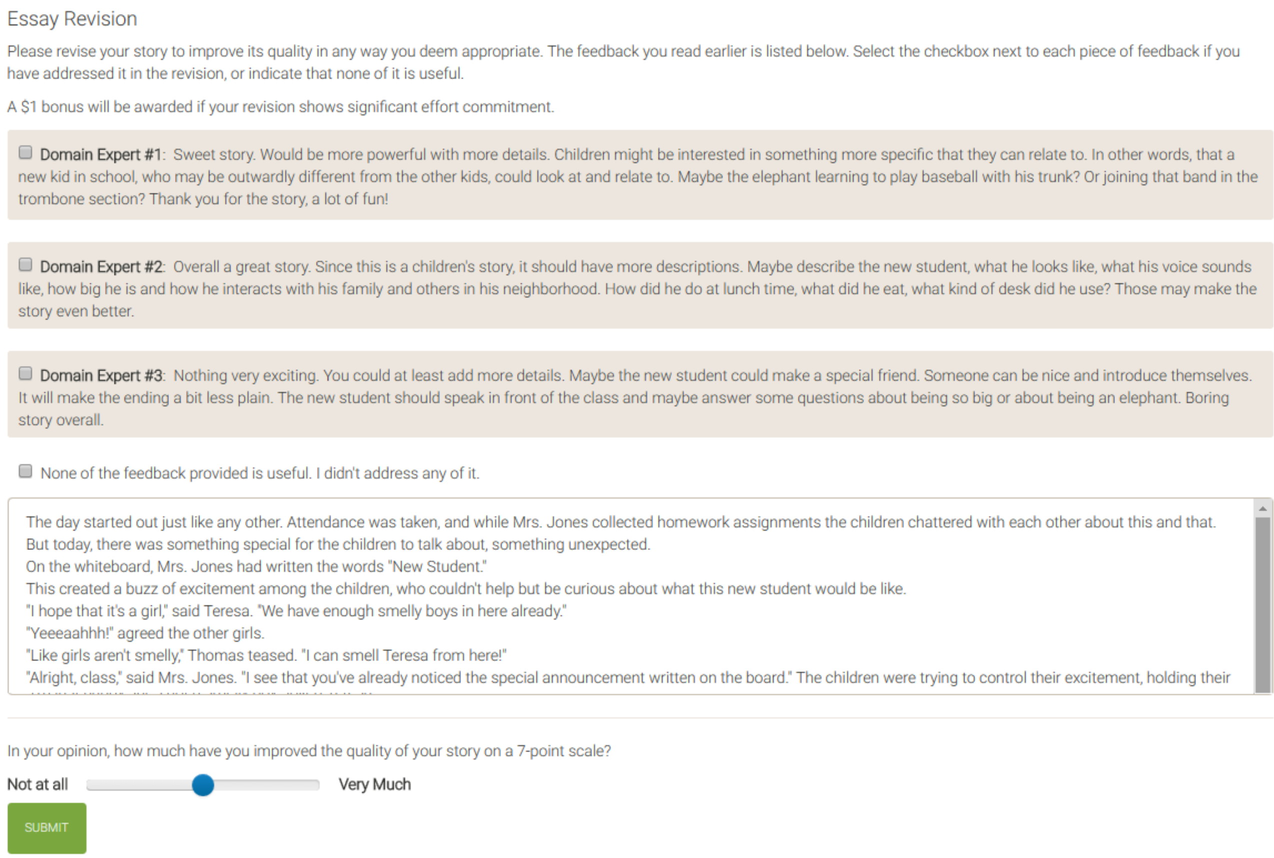

Figure 1. Task interface for the revision phase. Feedback was provided in different orders and with different source cues based on experimental conditions. Participants have already read the feedback piece by piece before reaching this stage.

peers are more responsive and provide more design suggestions in comparison with online design forums [42].

Besides expert and peer sources of feedback, in our work we also explored an anonymous condition, where no identity information about the source is given. Anonymity removes the social interaction element in the feedback interpretation process. Prior work shows the lack of social cues increases participants' motivation, perceived ability, and task performance when receiving computer-generated feedback [3]. Nguyen et al. show anonymity also increases feedback acceptance [24]. Our study extends this corpus of prior research by testing how feedback valence order interacts with different source identity cues.

METHODOLOGY

Our experiment addresses three research questions:

- How does ordering a feedback set based on valence affect the extent and quality of the subsequent content revisions?
- How does valence order influence a content creator's affective state, and influence his or her perceptions of the feedback and its source?
- How does information about the source of the feedback providers affect these same measures, and mediate the effects of the feedback ordering based on valence?

Answers to these questions help content creators better utilize feedback received online (e.g., positive affective state is associated with increased creative thinking [29]). Also, the answers can help platform designers know how to more effectively present the feedback (e.g., how to order it and whether to display source cues).

Experiment Design

To answer these research questions, we conducted a 3x3 full factorial online experiment with two factors: *Valence Order* and *Source*. Each participant received three pieces of feedback, including two with positive valence levels and one with negative valence level. There are three levels in *Valence Order*: negative first, negative between, and negative last. There are also three levels in *Source*: peer, expert, and anonymous. In the peer and expert conditions, the task instruction clearly states that the feedback comes from peer workers or domain experts. The feedback text also started with the words "Peer Worker" or "Domain Expert" (see Figure 1). In the anonymous condition, participants received

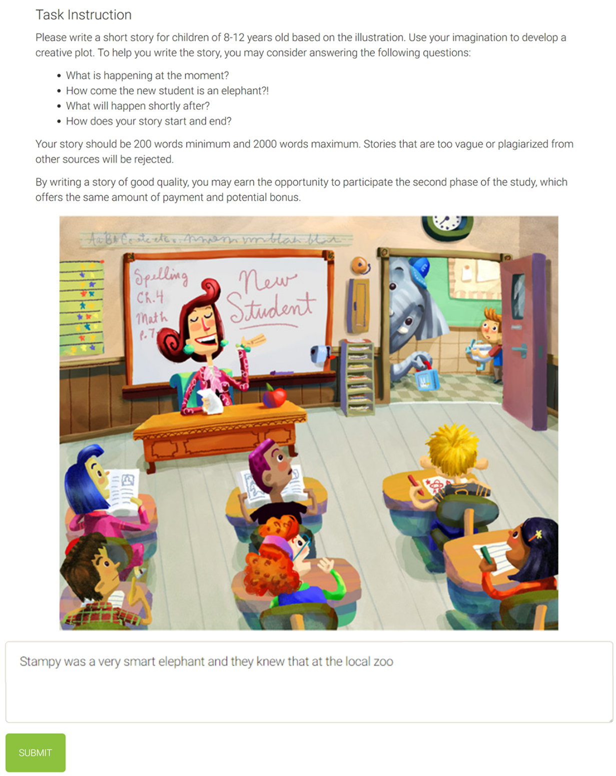

Figure 2. The user interface during the story writing phase. The image was edited for brevity.

no information about the feedback source. There was also no source identity cue in the feedback text.

Essay Task

The task includes two phases: *story writing* and *revision*. For the story writing phase, we asked participants to write a story for children of 8-12 years old based on a given illustration (Figure 2). A pilot study showed that most participants could finish the story in less than one hour. We intentionally allocated extra time for the task so the story quality would not be compromised due to time pressure. The participants had two hours to write a story within a 200-2000 word limit. We chose story writing as our experimental task for three reasons; (i) it is a topic that should be familiar to a general audience; (ii) it requires creative thought; and (iii) it only requires text entry, making it suitable to perform online.

The illustration facilitated the task in two ways. First, it provided scaffolding in the open-ended writing process by outlining the story's main characters and scenario. Second, it allowed the research team to select general feedback that applied to most stories by narrowing the scope of the possible story themes [34]. An expert in story writing was recruited and selected the illustration based on task appropriateness. Participants received $3 for the writing phase.

In the story revision phase, participants revised stories based on a set of feedback. After reviewing their stories at the start of the phase, participants received three pieces of feedback. The task interface presented one piece of feedback at a time. Participants selected a button to reveal the next piece of feedback. After feedback delivery, we asked participants to complete a survey about their affective states and perceptions of the feedback set and its perceived source. Then participants revised their original story to improve its quality. Participants received an additional $2 for the revision phase. To discourage satisficing, we offered a $1 bonus if they demonstrated significant effort during the revision phase. In total, top 30.9% of all participants ranked by edited character count received the bonus.

Feedback Pool

The feedback assigned to participants came from a feedback pool consisting of six pairs of positive and negative feedback (Table 1). Each pair was adapted from one piece of authentic feedback collected online. We used five stories from a pilot study to collect a large set of authentic feedback on the story plots. From the set, we selected six pieces of feedback that gave revision advice on story content. The feedback type was decided based on prior work about effective feedback [26].

To ensure each piece of feedback suggested a similar degree of revision, we recruited 30 online judges from Amazon Mechanical Turk to rate the actionability levels. Each judge reviewed 7 pieces of feedback including the 6 pieces of authentic feedback and one duplicate for quality control. For each piece of feedback, the judge rated the extent of revision needed if the feedback was accepted on a 7-point scale from 1 (No Revision Needed) to 7 (Major Revision Needed). We discarded the ratings from judges who rated the duplicate piece of feedback noticeably different (larger than two units) from its counterpart. The final average actionability rating across the feedback set was 4.12 (SE=0.26), and there was no significant difference between the ratings of any two pieces of feedback. The valence levels of the feedback pool were also adjusted and validated in the same manner. In the end, all adjusted positive feedback have similar positive valence levels (μ=5.38, SE=0.26), and the negative feedback have similar negative valence levels (μ=2.68, SE=0.20).

Participants

In total, 270 participants completed the experiment. All participants were recruited from Amazon Mechanical Turk and located in the U.S. All participants had finished more than 500 HITs and had a minimum 95% approval rate. Among the participants, 40% were males, 60% females; 98.1% reported English as their first language; 86.3% had read stories to children; 39.3% were parents of children younger than 15 years old. Regarding the age distribution, 9.6% were 18-24 years old, 42.2% 25-34 years old, 27.4% 35-44 years old, and 20.7% 45 years or older. Regarding the education level distribution, 40.8% reported high school or lower as their highest academic degree earned, 44.8% undergraduate degree, and 14.4% graduate degree. Since most workers on AMT earn at or below minimum wage [31], we assume that the participants were novice content creators rather than professional writers.

Positive Valence Version	Negative Valence Version
I really loved the story. I would change a couple of things. The story would be more interesting if more details were given. For example, what happened to his parents? In what ways does he feel different from the children? Does he miss living like an elephant? I teach young children and I would read this story to my class.	A pretty boring story. I would change a couple of things. The story would be more interesting if more details were given. For example, what happened to his parents? In what ways does he feel different from the children? Does he miss living like an elephant? I teach young children and I may not read this story to my class.
Great story! I think the new student also needs to introduce himself to the class so they can learn more about him. He can tell them where he is from, about how it is different from his new home area, what he likes to do, etc, so they can get to know him. The teacher can also ask the classmates to speak up if there is anything they like that the new student likes. That may make both the new student and the classmates more comfortable and willing to accept each other.	Quite a boring story. Some more details may make the story less plain. I think the new student also needs to introduce himself to the class so they can learn more about him. He can tell them where he is from, about how it is different from his new home area, what he likes to do, etc, so they can get to know him. The teacher can also ask the classmates to speak up if there is anything they like that the new student likes. That may make both the new student and the classmates more comfortable and willing to accept each other.
Overall a great story. Since this is a children's story, it should have more descriptions. Maybe describe the new student, what he looks like, what his voice sounds like, how big he is and how he interacts with his family and others in his neighborhood. How did he do at lunch time, what did he eat, what kind of desk did he use? Those may make the story even better.	Overall a pretty boring story. Since this is a children's story, it should at least have more descriptions. Maybe describe the new student, what he looks like, what his voice sounds like, how big he is and how he interacts with his family and others in his neighborhood. How did he do at lunch time, what did he eat, what kind of desk did he use? Those may make the story less boring.
Great story. You should add more details to make it even better. I would like for the new student to make a special friend. Maybe someone can be nice and introduce themselves. It will be an even happier ending to the story. The new student should speak in front of the class and maybe answer some questions about being so big or about being an elephant. Good job overall!	Nothing very exciting. You could at least add more details. Maybe the new student could make a special friend. Someone can be nice and introduce themselves. It will make the ending a bit less plain. The new student should speak in front of the class and maybe answer some questions about being so big or about being an elephant. Boring story overall.
Sweet story. Would be more powerful with more details. Children might be interested in something more specific that they can relate to. In other words, that a new kid in school, who may be outwardly different from the other kids, could look at and relate to. Maybe the elephant learning to play baseball with his trunk? Or joining that band in the trombone section? Thank you for the story, a lot of fun!	Boring story. Would be less plain with more details. Children might be interested in something more specific that they can relate to. In other words, that a new kid in school, who may be outwardly different from the other kids, could look at and relate to. Maybe the elephant learning to play baseball with his trunk? Or joining that band in the trombone section? I didn't really enjoy reading the story.
My daughter may love this story. I would like to read a bit more about the elephant's first day in the classroom - how he sat down, how the other children reacted, how he participated in the classroom work, and how the teacher treated him. I think that those details might add some more color to the story and perhaps even a bit more tension.	I probably wouldn't read this story to my daughter, just too boring. Maybe you can talk a bit more about the elephant's first day in the classroom - how he sat down, how the other children reacted, how he participated in the classroom work, and how the teacher treated him. I think that those details might add some more color to the story and perhaps even a bit more tension.

Table 1. The feedback pool from which the research team assigned three pieces of feedback to each initial story. At most one piece of feedback was assigned from each feedback pair (each row). The left and right columns show the positive and negative valence versions of the feedback, respectively.

Procedure

Participants read an IRB consent form and filled out a demographic survey at the beginning of the experiment. The task instruction also informed the participants about the following revision phase. Then we gave participants two hours to compose their initial stories. After all stories were collected, the research team selected three feedback pairs from the feedback pool for each story based on their appropriateness for the plot (Table 1). Each pair included one piece of positive and one piece of negative feedback, and both pieces were derived from the same piece of authentic feedback. For the three feedback pairs selected by the research team, a Python script randomly selected one piece of feedback from each pair, and two piece of positive feedback and one piece of negative feedback in total. For 22.3% of the stories, there were fewer than three feedback pairs applicable to the plot and we thus excluded these stories from the revision phase. While assigning feedback for all stories, we monitored how many times each feedback pair had been selected and adjusted to ensure an even allocation of the 6 pairs. Two days after participants finished the writing phase, we launched the revision phase and notified the participants via email. 74.6% of all qualified participants completed the revision phase. The task presented the feedback in different valence orders and with different source cues based on experiment conditions (Figure 1). The participants had two hours to finish the revision phase.

Measurements

We collected three sets of measurements:

- Affective states: how distressed / upset / enthusiastic / inspired / excited / happy they felt after receiving the feedback.
- Perceptions of the feedback and its source: how useful / positive / fair they perceived the feedback to be, how knowledgeable / polite the feedback sources to be, and how good they perceive their writing skill to be after receiving the feedback.
- Revision: how much time participants spent writing the initial story, reading the feedback, and revising the story; how much the story changed during the revision

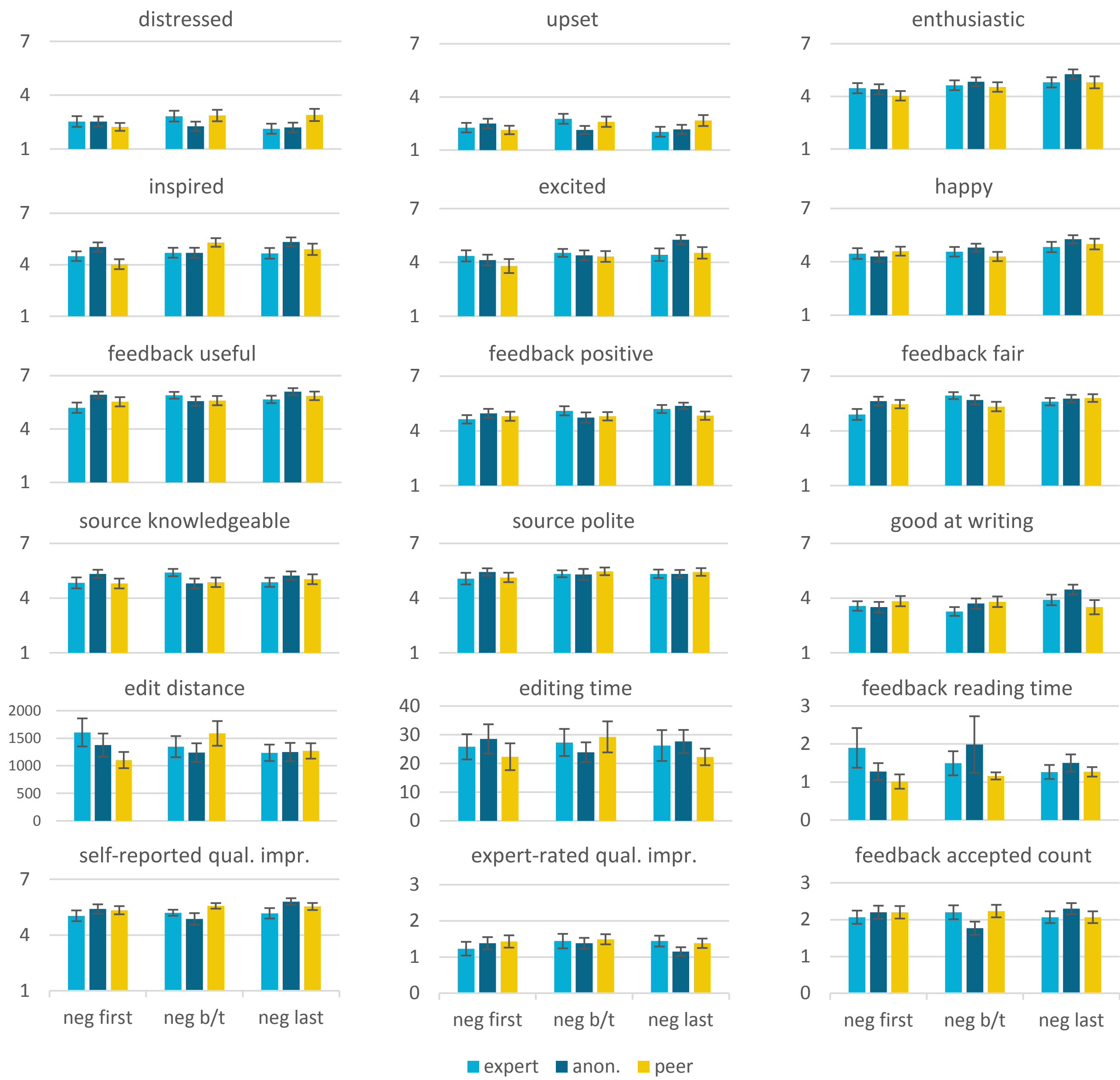

Figure 3. Measurements collected during the experiment. For all charts, the left / center / right groups of bars represent results from negative feedback presented first / between / last conditions. We also color-coded all bars according to source conditions: expert, anonymous (darkest), and peer (lightest). The vertical axes cover both the minimum and maximum (if applicable) range of the measurements.

in terms of self- and expert-rated quality improvement, and edit distance between the initial and revised stories.

We collected the first two sets of measurements from the survey, which included 13 statements regarding participants affective states (6 items adapted from PANAS [36]), their perceptions of the feedback (3 items) and its source (2 items), and confidence in writing skills (2 items). Metrics related to revision extents were derived from participants' action logs. For the quality improvement rating, two experts in English writing each rated all 270 stories. The rating interface presented both the initial and revised versions side by side, and the experts rated how much the revision had improved the quality of the story on a 7-point scale (-3: the original has much higher quality; +3: the revised has much higher quality; 0: no noticeable quality difference). For 91.7% of all stories, the rating difference between the experts was smaller or equal to one unit on the scale. We averaged the two ratings as the final quality improvement. The average quality improvement was 1.37 (SE=0.17) across all conditions.

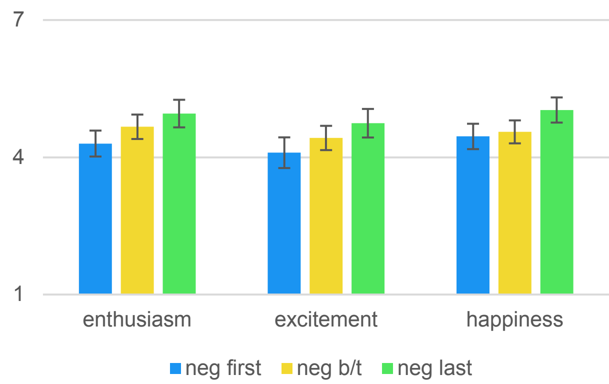

Figure 4. This chart shows the ratings for enthusiasm, excitement, and happiness; clustered by feedback ordering. Presenting negative feedback last resulted in higher ratings of participants' affective states.

RESULTS

In the following subsections, we report the most interesting patterns in our data. Figure 3 summarizes the results.

Participants were most motivated when reading negative feedback last

An ANOVA showed that *Valence Order* had a main effect on the ratings of enthusiasm ($F[2, 267]=3.96$, $p=.02$), excitement ($F[2, 267]=3.51$, $p=.03$), and happiness ($F[2, 267]=3.61$, $p=.03$). See Figure 4. Participants reported significant higher ratings in the negative last condition (enthusiasm: $\mu=4.96$, $SE=0.17$; excitement: $\mu=4.73$, $SE=0.18$; happiness: $\mu=5.01$, $SE=0.16$) than in the negative first condition (enthusiasm: $\mu=4.30$, $SE=0.16$; excitement: $\mu=4.10$, $SE=0.17$; happiness: $\mu=4.46$, $SE=0.16$; $p<.05$). Simply presenting the same feedback in different orders increased participants' enthusiasm by 11.0%, excitement by 10.5%, and happiness by 9.2% on a 7-point scale. In general, the later in the order participants read the negative feedback, the more enthusiastic, more excited, and happier they were. In the negative last condition, participants might view the first two pieces of positive feedback as an affirmation of the quality of their stories and become more resilient to the influence of the negative feedback. There was no significant effect of *Source* on participants' affective states.

Participants reported most favorable perception of the feedback when reading the negative feedback last

Participants in negative last condition rated the feedback set marginally fairer ($\mu=5.70$, $SE=0.14$) than in the negative first condition ($\mu=5.33$, $SE=0.14$; $p=.06$). The feedback positivity and usefulness ratings also show the same trends, but did not reach statistical significance. On average, participants rate the feedback fairer by 6.2%, more positive by 5.3%, and more useful by 5.2% on a 7-point scale in the negative last condition in comparison with the negative first condition. *Source* did not have a statistical effect on participants' perception of the feedback measured in the experiment.

Anonymity tended to improve affective states and perceptions of feedback and its source

Despite the lack of statistical significance, we observed interesting trends consistent with prior work regarding feedback source anonymity [24]. Our results show source anonymity tended to improve participants' affective states and led to more favorable perceptions of the feedback set and its source. In the anonymous condition, participants reported being 4.8% more enthusiastic, 5.7% more inspired, 4.6% more excited, 2.9% happier, 4.4% less distressed, and 2.6% less upset than in the peer and expert conditions. Participants also rated the feedback 4.1% more useful, 2.2% more positive, 3.4% fairer, and the feedback source 2.7% more knowledgeable, 1.4% politer on a 7-point scale in the anonymous condition.

Contrary to prior work [7], our results show there is no statistical difference in task performance between the expert and peer cue conditions. Participants in the expert condition spent more time reviewing the feedback ($\mu=94.3$ sec, $SE=12.2$) than in the peer condition ($\mu=68.8$ sec, $SE=12.2$), but they perceived the source in the peer condition ($\mu=4.90$, $SE=0.15$) to be nearly as knowledgeable as in the expert condition ($\mu=5.02$, $SE=0.15$). There is also no statistical difference in the edit distance or quality improvement.

One potential reason could be participants' familiarity with the writing task. 86.3% of the participants reported having told stories to young children. Participants with storytelling experience self-reported significantly higher quality improvement ($\mu=5.38$, $SE=0.08$) in comparison with the participants without ($\mu=4.89$, $SE=0.23$; $t(268)=2.17$, $p=.031$). Participants familiar with the task were more confident in their performance and their writing skill (participants w/ exp.: $\mu=3.78$, $SE=0.10$; w/o: $\mu=3.41$, $SE=0.26$). Prior work shows people with higher self-efficacy are less receptive to feedback [22]. The participants' familiarity with the task domain may therefore have affected how they perceived the source cues.

Improved affective states and feedback perception led to more revision

Neither *Valance Order* nor *Source* had a significant effect on the edit distance ($\mu=1327.56$, $SD=1025.82$), feedback accepted count ($\mu=2.12$, $SD=0.06$), and expert-rated quality improvement ($\mu=1.37$, $SD=0.86$). Edit distance was significantly correlated with the feedback accepted count (Pearson's $r=.41$; $p<.01$) and quality improvement ($r=.73$; $p<.01$). The more participants edited their essays, the higher the quality improvement ratings (Table 2). We also observed significant but weak correlations between the edit distance and ratings of enthusiasm ($r=.14$; $p<.05$), excitement ($r=.15$; $p<.05$), and inspiration ($r=.17$; $p<.01$). More motivated participants tend to revise their work more; therefore methods to improve motivation such as ordering feedback by valence as done in this study, providing immediate positive feedback [29], or wrapping feedback with positive language [24] as done in prior work can foster revision. Interestingly, participants' distress level also had a positive correlation

	happy	upset	distress	excited	enthusi.	insipired	fdbk pos.	fdbk use.	fdbk fair
edit distance	-0.03	0.10	0.13	0.15	0.14	0.17	0.12	0.18	0.21
accepted count	0.05	-0.02	-0.02	0.13	0.18	0.18	0.21	0.19	0.24
expert qual. impr.	-0.07	0.05	0.10	0.06	0.06	0.07	0.11	0.17	0.27

Table 2. Correlation table between revision extent metrics and participants' affective states and feedback perception. More favorable perception of the feedback set and more positive affective states correlate with a greater degree of revision.

with the edit distance (r=.13, p<.05). This correlation might be caused by the acceptance of negative feedback, which increased the edit distance and distress level at the same time. Participants were more likely to accept feedback that leaves more favorable impressions. Edit distance had significant correlations with how useful (r=.18, p<.01), positive (r=.12, p<.05), and fair (r=.21, p<.01) participants perceived the feedback set to be.

Female participants were more receptive to feedback

Table 3 shows the gender difference in our results. Prior work finds women are more influenced by verbal evaluative feedback than men [28]. In our experiment, female participants accepted significantly more pieces of feedback (μ=2.23, SE=0.07) and spent marginally longer time reading feedback (μ=1.62 min, SE=0.17) than male participants (μ=1.94, SE=0.09, t(268)=-2.53, p=.012; μ=1.16 min, SE=0.15, t(268)=-1.92, p=.056). Female participants also edited their stories more (μ=1495.35, SE=80.53), spent more time editing (μ=28.65, SE=1.93), and reported higher self-rated quality improvement (μ=5.49, SE=0.09) than male participants (μ=1075.87, SE=93.78, t(268)=-3.35, p<.001; μ=21.77, SE=2.32, t(268)=-2.27, p=.024; μ=5.03, SE=0.14, t(268)=-2.92, p=.004 respectively).

On the other hand, negative feedback had a stronger influence over female participants' affective states. Female participants reported feeling more distressed (μ=2.72, SE=0.12), more upset (μ=2.58, SE=0.12), and marginally less happy (μ=4.54, SE=0.14) compared to male participants (μ=2.17, SE=0.14, t(268)=-2.85, p=.005; μ=2.05, SE=0.13, t(268)=-2.89, p=.004; μ=4.88, SE=0.15, t(268)=1.86, p=.064). In sum, female participants were more likely to accept and be influenced by the feedback.

	Female	Male
accepted count *	2.23 (.07)	1.94 (.09)
feedback reading time	1.62 (.17) min	1.16 (.15) min
edited char count **	1495.35 (194.7)	1075.87 (147.7)
story editing time *	28.65 (1.93) min	21.77 (2.32) min
distressed **	2.72 (.13)	2.17 (.14)
upset **	2.58 (.12)	2.05 (.13)
happy	4.54 (.11)	4.88 (.15)

Table 3. Gender comparison between feedback receptivity and affective states. Standard errors of the means are included in parenthesis. For each row, '*' indicates significance level of p<.05, and '' indicates p<.01.**

DISCUSSION

Our results show presenting negative feedback last improves participants' affective states and perception of the feedback set. Cues of the feedback source had no significant effect on participants' affective states, and no effect on perceptions of the feedback set and its providers. There was no interaction between valence order and feedback source. Female participants were more likely to accept the feedback, even though reviewing the feedback had a larger negative impact on their affective states.

In our experiment, participants received three pieces of feedback for the initial story. This made the experiment tractable and gave the necessary control, but future work should test whether our results generalize to different sizes and valence balances of a feedback set. For larger feedback sets, platform designers could choose to select representative pieces of the feedback to summarize the larger set. Some online platforms, such as Amazon.com, have already adopted this method by showing the highest rated positive and negative reviews as a summary. This presentation mechanism allows users to quickly grasp the key insights without spending significant time consuming all reviews. Online design communities may explore similar techniques based on the valence level and the popularity of the feedback.

Similar techniques could also be used in creativity support tools for writing. When presenting comments collected from external reviewers, writing support tools could offer positive valence comments first and negative ones last. In the case where there is a large quantity of comments or when it is difficult to re-order the feedback (e.g., for inline comments), the tool could show only the positive comments as the default and users could access the additional feedback through interaction. On the other hand, tools could prompt feedback providers to write separately about the positive aspects of the work, and display a summary of these responses first. Future work could also test data-driven approaches that automate positive valence feedback. The system could compare content creators' performance, in terms of grammatical error rate or estimated vocabulary size, against their own prior writing or their peers, and report the positive results.

In our experiment, we achieved different levels of valence by adjusting the language of the feedback. Our results may also generalize to other visual indicators of valence. Platform designers in creative domains could take inspiration from other online review services. Some online work platforms deliver feedback along with valence indicators such as upvote / downvote in performance review or job approval / rejection scenarios. These indicators make it straightforward for platform designers to order feedback by valence. Another common form of a valence indicator is a numeric rating such as star ratings or scores on review sites. Fine-grained ratings make it easy to compare the valence levels among feedback. Platform designers could implement these valence indicators in online feedback collection services and facilitate the valence ordering process.

Feedback valence order had the same influence across the source conditions in our study. The valence order may therefore have similar effects on platforms where feedback providers have different social identities and expertise. The participants in our study were mainly novice content creators. Experienced content creators may react differently to the manipulations. Prior work shows experts seek negative feedback more actively than novices [10]. Negative feedback may therefore have a weaker impact on experienced content creators. Future work should test whether experience level of content creators interacts with valence order.

In our study, there was a delay between writing the initial story and receiving the feedback. For future work, it would be interesting to test how a delay between reviewing the feedback and revising the story would affect our measures, as delays can promote reflection and learning [14]. Another way to extend our work is to test the effects of different positive to negative feedback ratios, especially when a majority of the feedback has a negative valence. Researchers can also explore effects of ordering feedback using attributes other than valence such as design aspects [26], content complexity [6], and the effort invested in writing it [38].

Future work could also examine how *writing* feedback in a certain valence order affects the providers' perceptions of the usefulness of their own feedback. This technique could potentially improve feedback providers' experience on the review platform and encourage further participation.

CONCLUSION
Content creators frequently receive negative feedback online. To increase feedback receptivity, we tested a novel method of ordering a set of feedback based on its valence, and tested how this ordering interacts with different source identities. Our work makes two contributions. First, we provided empirical evidence that presenting negative feedback last improves content creators' affective states and their perception of the feedback. Second, our results show valence order does not interact with feedback source, although source anonymity tends to improve feedback acceptance. In general, our study demonstrates that a simple feedback order manipulation can lead to measurable improvements in content creators' affective states and perception of feedback. We hope our work can improve content creators' experience of receiving feedback online.

ACKNOWLEDGMENTS
This work was supported in part by the National Science Foundation under awards NSF CMMI 14-62693 and IIS 15-30818.

REFERENCE

1. Matthijs Baas, Carsten K W De Dreu, and Bernard A Nijstad. 2008. A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin* 134, 6: 779–806. https://doi.org/10.1037/a0012815
2. Jessica L. Bienstock, Nadine T. Katz, Susan M. Cox, Nancy Hueppchen, Sonya Erickson, and Elizabeth E. Puscheck. 2007. To the point: medical education reviews-providing feedback. *American Journal of Obstetrics and Gynecology* 196, 6: 508–513. https://doi.org/10.1016/j.ajog.2006.08.021
3. Cheryl Campanella Bracken, Leo W. Jeffres, and Kimberly A. Neuendorf. 2004. Criticism or Praise? The Impact of Verbal versus Text-Only Computer Feedback on Social Presence, Intrinsic Motivation, and Recall. *CyberPsychology & Behavior* 7, 3: 349–357. https://doi.org/10.1089/1094931041291358
4. Peter Cantillon and Joan Sargeant. 2008. Giving feedback in clinical settings. *BMJ (Clinical research ed.)* 337, 7681: a1961. https://doi.org/10.1136/bmj.a1961
5. Olle Th J ten Cate. 2013. Why receiving feedback collides with self determination. *Advances in Health Sciences Education* 18, 4: 845–849. https://doi.org/10.1007/s10459-012-9401-0
6. John V Dempsey, Marcy P Driscoll, and Linda K Swindell. 1993. Text-based feedback. *Interactive instruction and feedback*: 21–54.
7. D Fedor. 2001. Performance improvement efforts in response to negative feedback: the roles of source power and recipient self-esteem. *Journal of Management* 27, 1: 79–97. https://doi.org/10.1016/S0149-2063(00)00087-8
8. Dana R Ferris. 1997. The Influence of Teacher Commentary on Student Revision. *TESOL Quarterly* 31, 2: 315. https://doi.org/10.2307/3588049
9. Dana Ferris and Barrie Roberts. 2001. Error feedback in L2 writing classes. *Journal of Second Language Writing* 10, 3: 161–184. https://doi.org/10.1016/S1060-3743(01)00039-X
10. Stacey R. Finkelstein and Ayelet Fishbach. 2012. Tell Me What I Did Wrong: Experts Seek and Respond to Negative Feedback. *Journal of Consumer Research* 39, 1: 22–38. https://doi.org/10.1086/661934
11. Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition - C&C '15*, 235–244. https://doi.org/10.1145/2757226.2757249
12. Laura Hale. 2015. What is the meaning of "answer collapsed" in Quora? (August 22, 2015). *Quora*. Retrieved January 5, 2017 from

https://www.quora.com/What-is-the-meaning-of-answer-collapsed-in-Quora/answer/Laura-Hale?srid=uxyk

13. Richard H. Haswell. 2005. NCTE/CCCC's recent war on scholarship. *Written Communication* 22, 2: 198–223. https://doi.org/10.1177/0741088305275367

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

15. John Hedgcock and Natalie Lefkowitz. 1994. Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing* 3, 2: 141–163. https://doi.org/10.1016/1060-3743(94)90012-4

16. E a Hesketh and J M Laidlaw. 2002. Developing the teaching instinct. *Medical teacher* 24, 3: 245–348. https://doi.org/10.1080/014215902201409911

17. David G. Jansson and Steven M. Smith. 1991. Design fixation. *Design Studies* 12, 1: 3–11. https://doi.org/10.1016/0142-694X(91)90003-F

18. Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to participate in online communities. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 1927. https://doi.org/10.1145/1753326.1753616

19. Kurt Luther, Jari-lee Tolentino, Wei Wu, et al. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 473–485. https://doi.org/10.1145/2675133.2675283

20. Winter Mason and Duncan J. Watts. 2010. Financial incentives and the "performance of crowds." *ACM SIGKDD Explorations Newsletter* 11, 2: 100. https://doi.org/10.1145/1809400.1809422

21. Edna Holland Mory. 2004. Feedback Research Revisited. In *Handbook of research on educational communications and technology* (2nd ed.), David H. Jonassen (ed.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 745–783.

22. AnJanette A. Nease, Brad O. Mudgett, and Miguel A. Quiñones. 1999. Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *Journal of Applied Psychology* 84, 5: 806–814. https://doi.org/10.1037/0021-9010.84.5.806

23. Christine M Neuwirth, Ravinder Chandhok, Davida Charney, Patricia Wojahn, and Loel Kim. 1994. Distributed collaborative writing. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*, 51–57. https://doi.org/10.1145/191666.191693

24. Thi Thao Duyen T. Nguyen, Thomas Garncarz, Felicia Ng, Laura A. Dabbish, and Steven P. Dow. 2017. Fruitful Feedback. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1024–1034. https://doi.org/10.1145/2998181.2998319

25. Jay Parkes, Sara Abercrombie, and Teresita McCarty. 2013. Feedback sandwiches affect perceptions but not performance. *Advances in Health Sciences Education* 18, 3: 397–407. https://doi.org/10.1007/s10459-012-9377-9

26. J. van der Pol, B.A.M. van den Berg, W.F. Admiraal, and P.R.J. Simons. 2008. The nature, reception, and use of online peer feedback in higher education. *Computers & Education* 51, 4: 1804–1817. https://doi.org/10.1016/j.compedu.2008.06.001

27. Mark C. Porte, George Xeroulis, Richard K. Reznick, and Adam Dubrowski. 2007. Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. *The American Journal of Surgery* 193, 1: 105–110. https://doi.org/10.1016/j.amjsurg.2006.03.016

28. Tomi-Ann Roberts and Susan Nolen-Hoeksema. 1989. Sex differences in reactions to evaluative feedback. *Sex Roles* 21, 11–12: 725–747. https://doi.org/10.1007/BF00289805

29. Alwin de Rooij, Philip J Corr, and Sara Jones. 2015. Emotion and Creativity: Hacking into Cognitive Appraisal Processes to Augment Creative Ideation. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition - C&C '15*, 265–274. https://doi.org/10.1145/2757226.2757227

30. Alwin de Rooij and Sara Jones. 2013. Mood and creativity. In *Proceedings of the 9th ACM Conference on Creativity & Cognition - C&C '13*, 362. https://doi.org/10.1145/2466627.2466658

31. Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, 2863. https://doi.org/10.1145/1753846.1753873

32. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2: 119–144. https://doi.org/10.1007/BF00117714

33. Jason Shah. 2012. Hacker News: Stop greying out flagged comments (March 10, 2012). Retrieved January 5, 2017 from http://blog.jasonshah.org/post/19079124354/hacker-news-stop-greying-out-flagged-comments

34. Steven M Smith, Julie S. Linsey, and Andruid Kerne. 2011. Using Evolved Analogies to Overcome Creative Design Fixation. In *Design Creativity 2010*. Springer London, London, 35–39. https://doi.org/10.1007/978-0-85729-224-7_6

35. John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3: 321–326. https://doi.org/10.1089/1094931041291295

36. D Watson, L a Clark, and A Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6: 1063–70. https://doi.org/10.1037/0022-3514.54.6.1063

37. Daniel B Willingham. 1990. Effective feedback on written assignments. *Teaching of Psychology* 17, 1: 10–13. https://doi.org/10.1207/s15328023top1701_2

38. Y Wayne Wu and Brian P Bailey. 2016. Novices Who Focused or Experts Who Didn't? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 4086–4097. https://doi.org/10.1145/2858036.2858330

39. Anbang Xu and Brian P. Bailey. 2012. What Do You Think? A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. *Proceedings of the 15th ACM conference on Computer supported cooperative work & social computing - CSCW'12*: 295. https://doi.org/10.1145/2145204.2145252

40. Anbang Xu, Shih-wen Huang, and Brian Bailey. 2014. Voyantt: 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*, 1433–1444. https://doi.org/10.1145/2531602.2531604

41. Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong. 2011. Review spotlight. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 1541. https://doi.org/10.1145/1978942.1979167

42. Yu-Chun Grace Yen, Steven P. Dow, Elizabeth Gerber, and Brian P. Bailey. 2016. Social Network , Web Forum , or Task Market ? Comparing Different Crowd Genres for Design Feedback Exchange. *ACM Designing Interactive Systems (DIS 2016)*: 773–784. https://doi.org/10.1145/2901790.2901820

43. Yu-Chun Grace Yen, Steven P. Dow, Elizabeth Gerber, and Brian P. Bailey. 2017. Listen to Others, Listen to Yourself: Combining Feedback Review and Reflection to Improve Iterative Design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition - C&C '17*. https://doi.org/10.1145/3059454.3059468

44. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Björn 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*, 1003–1015. https://doi.org/10.1145/2818048.2819953

45. Jing Zhou. 1998. Feedback valence, feedback style, task autonomy, and achievement orientation: Interactive effects on creative performance. *Journal of Applied Psychology* 83, 2: 261–276. https://doi.org/10.1037/0021-9010.83.2.261

46. Haiyi Zhu, Robert E Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 3431. https://doi.org/10.1145/1978942.1979453

47. Hiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. 2013. Effects of peer feedback on contribution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 2253. https://doi.org/10.1145/2470654.2481311

48. Sentiment Analysis API| AlchemyAPI. Retrieved January 5, 2017 from http://www.alchemyapi.com/products/alchemylanguage/sentiment-analysis