ABSTRACT
Paid crowdsourcing platforms, which harness the extrinsic motivation of monetary compensation, are seeing increased usage for design feedback tasks, but the low quality of design feedback from crowdsourced workers continues to be a problem for designers. Intrinsic motivation has the potential to increase the quality of worker responses, but is difficult to elicit in paid workers. In this paper, we present the results of a study performed on the paid crowdsourcing platform Mechanical Turk in which we investigate the effect of introducing intrinsic motivation on the quality and character of workers’ responses to a design feedback task. Workers were assigned to either an intrinsic condition, where they were allowed to choose the topic of the design they would give feedback on, or an extrinsic condition where they were assigned a topic randomly. Analysis of results showed that giving workers a choice of topic did not affect the quality of their responses but did cause them to write more, more efficiently, and more confidently than workers in the extrinsic condition.

RELATED WORK
Intrinsic motivation is an often overlooked feature of high quality work. Success in graduate school is hugely dependent on whether or not that student has the motivation to work independently for years at a time on a difficult problem. Video games, sports, and learning new hobbies can be very challenging, but those who are love the experience enough and want to get better persevere, many without even getting paid for it! If intrinsic motivation could be harnessed in a paid crowdsourcing platform like Mechanical Turk, it could go a long way towards improving the quality of responses without needing to pay more, recruit lots of workers, or build additional software.

Using intrinsic motivation on a paid crowdsourcing platform like Mechanical Turk is not as far-fetched as it sounds. A 2010 study found that 59% of Indian workers and 69% of US workers agree that “Mechanical Turk is a fruitful way to spend free time and get some cash” [2]. Thus, while most workers find payment relevant, most also consider the money they make as something extra and not necessary. Given that increasing payment does not increase response quality on Mechanical Turk[6], maybe it’s time to try intrinsic motivation.

Inevitably, any attempt to study intrinsic motivation on a paid platform will inevitably mix intrinsic and extrinsic motivation together, and the two often interact in unexpected ways. In an experiment performed by Deci, it was found that students paid to play with a puzzle later played with it less and reported less
We aim to investigate whether intrinsic motivation can be used to improve the quality, or at least the character, of design feedback on a paid crowdsourcing platform. We also want to know what the effects of a complex interaction between intrinsic and extrinsic motivation may be for an online design feedback task.

**EXPERIMENTAL DESIGN**

To measure the effects of extrinsic and intrinsic motivation on Mechanical Turk, we created a visual design feedback task with two conditions: Personal and Random. In both conditions, a worker received one of 13 different website design screenshots to critique, but the manner of choosing a specific design was what differed between the conditions. In the Personal condition, workers were asked to choose one of 13 topics that they found the most interesting, and then were given a webpage related to the chosen topic. In the Random version, workers were assigned a topic randomly. The intuition behind these two conditions is that workers given a choice of topic will have a higher personal investment in the task, and thus a more intrinsic motivation, than those that are simply assigned a random design. Since AMT workers are typically wary of clicking URLs to external websites and we needed to keep what workers were critiquing consistent across conditions and between individual workers, we showed workers a screenshot of each website’s main page instead of giving them a link to the website.

The 13 topics we chose for this experiment can be seen in Table 1. We chose a set of topics that had broad enough appeal to ensure that at least one of them would interest any given worker. In addition to asking for a critique of the webpage design, workers were also asked how much money they felt the task was worth in order to determine how intrinsic versus extrinsic motivation affected the perceived value of the task.

**PILOT STUDY**

Before launching a full study with the Personal and Random conditions, we launched a smaller scale study on Mechanical Turk with 5 HIT assignments per condition (10 in total) to make sure our HITs were ready. Each worker was paid $0.30, each worker could only complete one HIT assignment in each group, and we restricted workers to the US because many of the topics we chose for website designs concerned hobbies that were very American-centric (for example, Hiking, Pets). After presenting the chosen design (chosen by the worker in the Personal condition or chosen randomly in the Random condition), we asked workers in both conditions the following questions:

2. How could this website design be improved? Provide some specific, actionable suggestions.
3. You are being paid $0.30 for this task. In your opinion, how much is this task worth? If we were to post it again, how much should we pay?

We chose the adjectives in Question 1 based on the adjectives used for describing visual aesthetics in a website design described in Lavie et al’s paper[3]. These adjectives cover the broader categories of classical aesthetics and expressive aesthetics, and we believed they would help novice workers give more specific, targeted design critiques. Question 2 was also designed to get workers to elicit more specific responses and avoid generic, vague statements.

After all HIT assignments were completed, all four of our group members rated the quality of the responses to Question 1 and Question 2 independently on a 7-point Likert scale. The results of the pilot study indicated that workers in the Personal condition spent more time on the task (394 seconds versus 223 seconds in the Random condition), had higher average quality ratings for Questions 1 and 2 (5.05 and 4.98 in the Personal condition versus 4.85 and 4.78 for the Random condition, respectively), wrote more in their critiques (54.5 words versus 34.8 words in the Random condition), and, surprisingly, expected a higher reward than workers in the Random condition ($0.54 versus $0.44 per task). This last result is surprising given our expectation that workers in the Personal condition would be less motivated by money and thus more willing to accept a lower reward.

After analyzing the pilot study results, we realized that we needed to make four main changes to our experimental design:

1. The specific designs we chose for each topic varied quite a bit in their visual complexity, which may have introduced unnecessary variance into our results.
2. Our two critique questions were too complicated and differences in motivation between our two conditions might have been masked by the very specific prompts.
3. We had no way of assessing whether the specific design workers received was actually interesting to a worker, regardless of whether they had chosen their own topic or not.
4. None of us were blind to condition when rating response quality, and this may have affected our quality ratings.

**MAIN EXPERIMENT**

We modified our HITs for the Personal and Random conditions to fix the problems we observed in the pilot study. In order to make sure all of the website designs we chose for our experiment were visually consistent, we chose one topic’s design as our “gold standard” for visual complexity. We decided that the Allrecipes.com website used for our Cooking topic fell somewhere in the middle of the visual complexity scale, with an even balance of images and text (Fig. 1). For the other 12 topics, we found 2 additional...
Table 1. Selected topics for the experiment

<table>
<thead>
<tr>
<th>Art; Cooking</th>
<th>Fitness</th>
<th>Hiking</th>
<th>Movie</th>
<th>Music</th>
<th>Pets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photography</td>
<td>Reading</td>
<td>Shopping</td>
<td>Sport new</td>
<td>Tech</td>
<td>Video games</td>
</tr>
</tbody>
</table>

Figure 1. Allrecipes.com, our standard for visual design complexity. We selected a website for each topic based on their visual similarity to this website.

Figure 2. Topic distribution in Personal and Random conditions

RESULTS

We performed a pilot study on 10 participants (5 each condition) and a main experiment on 60 participants (30 each condition). The results of the main experiment are analyzed in this section. For a discussion of the pilot study results, see the Experimental Design section.

For main experiment, we analyzed the topic distribution, quality rating, task completion time, word count, sentiment, content similarity, expected amount of reward, and workers’ interest score based on the feedback we got in the Random and Personal conditions. We conduct ANOVAs for quality rating, task completion time, word count, expected amount of reward, and workers’ interest score between the Personal and Random conditions.

We also investigate whether workers’ reported interest rating influenced their responses independently of what condition they were assigned to. We separate workers’ interest ratings into two buckets for analysis: interest ratings greater than 3 and interest ratings less than or equal to 3. The ANOVA results showed that interest score had a significant influence on word count and sentiment. People tend to give longer and more positive feedback when they are more interested (interest rating > 3) in the topic. For the other features, there is no significant difference between the two interest rating groups.

Topic distribution

In the Personal condition, workers can choose their own website topic for their critiques. In the Random condition, topics are assigned to workers randomly. Fig. 2 shows the topic distribution for both conditions. In the Personal condition, the Art, Video Games, and Cooking websites are chosen the most, while the Reading and Photography websites were not chosen by any worker. We conduct Fisher’s Exact Test for Count Data on the topic distributions since many topics were chosen fewer than 5 times. We find no significant difference (p-value=0.487) between the distributions of the two conditions, suggesting that there was not much bias for particular topics in the Personal condition.
Quality Rating
In the main experiments, three members in our team, blind to condition, rated the quality of feedback on a 7-point Likert scale. The mean quality rating in the Personal condition was 3.80, while in the Random condition, it was 4.01. Fig. 3 shows a boxplot of mean quality ratings. There was no significant difference (p-value=0.448) between the two conditions’ quality ratings.

The mean quality ratings for the high interest and low interest groups were 3.6 and 4.0, respectively. There was also no significant difference between these two groups (p-value=0.23).

Task Completion Time
The mean task completion time in the Personal condition was 282.6 seconds, and in the Random condition, 334.7 seconds. There is no significant difference (p-value=0.31) between the Personal and Random condition.

The mean task completion times for the low interest group and the high interest group were 237.9 and 326.5 seconds, respectively. When we conducted an ANOVA on the two groups, we saw no significant influence difference (p-value=0.18) on task completion time.

Word Count
The mean word count was 82.4 in the Personal condition and 60.9 in the Random condition. As shown in Fig. 4, there is a significant difference between conditions for word count (p-value = 0.034). People tend to write more feedback when they are allowed to select website design by themselves.

The mean word count in the low interest group was 56.0, while in the high interest group, it was 77.0 (Fig. 5). There was no significant difference between the two groups, although the p-value approaches significance (p-value=0.11), suggesting that people might write more when they are more interested in the topic of the website design, regardless of whether the topic was assigned randomly or not.
Sentiment

We used Python NLTK to analyze the sentiment of workers’ critique responses. The positive sentiment scores in the Personal and Random conditions were 0.42 and 0.51, respectively. The ANOVA showed no significant difference in sentiment (p-value=0.43) between the two conditions.

The mean sentiments for the low interest group and high interest group are 0.42 and 0.51, respectively, and interest rating has an almost significant positive influence (p-value=0.12) on the sentiment of feedback.

Content Similarity

We performed simple K-Means clustering using the Weka machine learning software package, with the Euclidean distance as the default metric. The K-means algorithm reported 2 clusters: Cluster-0 with 56 responses and Cluster-1 with only four responses (Fig. 6). Cluster-0 contained 29 Personal condition responses and 27 Random condition responses. Cluster-1 contained 3 Personal condition responses and 1 Random condition response. The accuracy of the clustered data was 53.34%. The resulting clusters imply that the content of Personal and Random condition responses do not significantly differ.

We made two word clouds (Fig. 7 and 8) for feedback in the Personal and Random condition. Twelve out of thirty people in the Random condition start with praise like “I like the...” “I really enjoy...” “It’s a beautiful website...”. From the word clouds, we also found some differences in key words. In the Personal condition, people mentioned “background” and “art” more often. This may be because two topics, Art and Video Games, were selected by many people in the Personal condition, and “background” and “art” are highly related to design critique for these two topics. Moreover, we notice that the words “perhaps” and “maybe” appear many more times in the Random condition, while “should” appears many more times in the Personal condition.

Expected Reward

Fig. 9 shows the expected reward histogram for the Personal and Random conditions. The means for expected reward in the Personal and Random conditions were $0.74 and $0.41, respectively. The mean of the Personal condition was likely skewed high by the single person who asked for $10.00. If we consider this response an outlier and perform an ANOVA that excludes it, then there is no significant difference in expected reward for the Personal and Random groups (p-value = 0.55).
We found that allowing workers to choose their own topic. Workers could be interested in their selected topic, but not necessarily. We did sentiment analysis to determine whether allowing workers to choose their own topic would affect the sentiment of critiques. The sentiment in the Personal and Random conditions is quite similar. We did not observe a significant difference between low and high interest groups in sentiment. There were no significant differences between the Personal and Random conditions for task completion time, and interest rating also did not affect task completion time. This is an especially puzzling result given that word count was significantly higher for the high interest group. Perhaps workers who are intrinsically motivated may work more efficiently on a task than those with only extrinsic motivation.

We observed little content-related difference between the Personal and Random condition when using k-means clustering, but manually examining the content led to some interesting results. People tend to give suggestions directly in the Personal condition while in the Random condition, many people give praise first and then suggestions. Perhaps those who chose the topic themselves (Personal condition) felt their opinions would be taken more seriously and thus did not feel the need to give praise first. Or maybe since they knew more about the chosen topic, Personal condition workers noticed the problems first and then the good things about the website afterwards. In the Personal condition, people mentioned “background” and “art” more often, perhaps because the two most popular topics in the Personal condition were Art and Video Games, and “background” and “art” are highly related to design critique for these two topics.

In addition, based on the higher frequency of assertive language in the Personal condition compared to the Random condition, it seems people in the Personal condition are more confident in giving feedback. We think this may be because they are more familiar with the topic of the website, having chosen it themselves.

There was no effect of condition on workers’ expected reward. This is consistent with the lack of significant differences between the conditions on most other metrics and is likely indicative that allowing workers to choose their own topic did not provide enough of an intrinsic motivation for workers to behave significantly differently in the Personal and Random conditions, or even in the high interest and low interest groups.

It seems that we simply need a stronger intrinsic incentive in order to see differences between intrinsically and extrinsically motivated workers in a controlled setting.

The interest rating results reveal another potential problem with our methodology. Since so many workers in the Random condition gave the highest possible interest rating, many of them likely answered dishonestly, perhaps out of fear that their results would be rejected if they rated the task not as interesting. Or perhaps these workers’ expectations were lower going into the task because they did not know what sort of design they would be getting to critique. Either way, if we want to have honest self-reported interest ratings for future studies, it seems we must think of a way to pose the question to workers in a way that does not imply that their response will affect their payment.

DISCUSSION

We find that people tend to write more feedback when they’re allowed to select the topic of the website they will critique. This may because people are more familiar and interested in the topic they select and thus would have more comments as well as give more suggestions. Therefore, if we want to get longer feedback, we need to recruit people who are interested in the related topics.

We did sentiment analysis to determine whether allowing workers to choose their own topic would affect the sentiment of critiques. The sentiment in the Personal and Random conditions is quite similar. We did not observe a significant difference between low and high interest groups in sentiment. There were no significant differences between the Personal and Random conditions for task completion time, and interest rating also did not affect task completion time. This is an especially puzzling result given that word count was significantly higher for the high interest group. Perhaps workers who are intrinsically motivated may work more efficiently on a task than those with only extrinsic motivation.

We observed little content-related difference between the Personal and Random condition when using k-means clustering, but manually examining the content led to some interesting results. People tend to give suggestions directly in the Personal condition while in the Random condition, many people give praise first and then suggestions. Perhaps those who chose the topic themselves (Personal condition) felt their opinions would be taken more seriously and thus did not feel the need to give praise first. Or maybe since they knew more about the chosen topic, Personal condition workers noticed the problems first and then the good things about the website afterwards. In the Personal condition, people mentioned “background” and “art” more often, perhaps because the two most popular topics in the Personal condition were Art and Video Games, and “background” and “art” are highly related to design critique for these two topics.

In addition, based on the higher frequency of assertive language in the Personal condition compared to the Random condition, it seems people in the Personal condition are more confident in giving feedback. We think this may be because they are more familiar with the topic of the website, having chosen it themselves.

There was no effect of condition on workers’ expected reward. This is consistent with the lack of significant differences between the conditions on most other metrics and is likely indicative that allowing workers to choose their own topic did not provide enough of an intrinsic motivation for workers to behave significantly differently in the Personal and Random conditions, or even in the high interest and low interest groups.

It seems that we simply need a stronger intrinsic incentive in order to see differences between intrinsically and extrinsically motivated workers in a controlled setting.

The interest rating results reveal another potential problem with our methodology. Since so many workers in the Random condition gave the highest possible interest rating, many of them likely answered dishonestly, perhaps out of fear that their results would be rejected if they rated the task not as interesting. Or perhaps these workers’ expectations were lower going into the task because they did not know what sort of design they would be getting to critique. Either way, if we want to have honest self-reported interest ratings for future studies, it seems we must think of a way to pose the question to workers in a way that does not imply that their response will affect their payment.

CONCLUSION

We set out to determine how extrinsic and intrinsic motivation affect the quality and type of design feedback on an online crowdsourcing platform both independently and in combination. We launched a controlled study on Mechanical Turk where workers were randomly assigned to either an intrinsic motivation condition or an extrinsic motivation condition. All
Table 2. Selected websites per topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art</td>
<td>Behance.com, Carbonmade.com, Art-deviantart.com</td>
</tr>
<tr>
<td>Fitness</td>
<td>Bodybuilding.com, Bornfitness.com, Fitbottomedgirl.com</td>
</tr>
<tr>
<td>Hiking</td>
<td>Alltrails.com, AmericanHikingSociety.com, Traillink.com</td>
</tr>
<tr>
<td>Movie</td>
<td>YahooMovies.com, MoviesDotCom.com, RottenTomatoes.com</td>
</tr>
<tr>
<td>Music</td>
<td>Di.com, Gaana.com, Jango.com</td>
</tr>
<tr>
<td>Pets</td>
<td>Petco.com, Petsmart.com, Petsupermarket.com</td>
</tr>
<tr>
<td>Photography</td>
<td>500px.com, Flickr.com, Polaroidblipfoto.com</td>
</tr>
<tr>
<td>Reading</td>
<td>Bookcoverarchive.com, Goodreads.com, Shelfari.com</td>
</tr>
<tr>
<td>Shopping</td>
<td>Bonadrag.com, Fab.com, Net-a-porter.com</td>
</tr>
<tr>
<td>Sport news</td>
<td>Fox.com, Sports.yahoo.com, ESPN.com</td>
</tr>
<tr>
<td>Tech</td>
<td>Cnet.com, Techcrunch.com, Techverge.com</td>
</tr>
<tr>
<td>Video games</td>
<td>Gamespot.com, Gamesradar.com, Ign.com</td>
</tr>
</tbody>
</table>

workers were asked to critique the visual design of a website, but in the intrinsic condition (Personal), workers were allowed to choose the topic of the website they would critique, whereas workers in the extrinsic condition (Random) were assigned a website from a random topic. We hoped to see whether workers’ responses in website design critique task would differ in terms of quality, task completion time, word count, content, sentiment, and the amount of money they believed they deserved to be paid for the task.

There were surprisingly few differences between conditions in the end. People assigned to the Personal condition tended to write more confidently, present criticisms before praise, and write more in the same amount of time than the Random condition did. These results indicate to us that for certain tasks where efficiency is important, it is necessary to recruit more intrinsically motivated individuals. However, the Personal and Random conditions had no other significant differences, and we observed no additional significant differences even when we separated our data by interest rating instead of condition. This suggests to us that the structure of our HITs and the tasks we assigned the workers to do were not conducive to promoting a distinction between extrinsic and intrinsic motivation. Future work will require us to redesign HITs and tasks to ensure that our intrinsic incentives actually work. And perhaps Mechanical Turk workers have too much extrinsic motivation for small intrinsic incentives to have a significant effect on their performance. Maybe it is necessary to investigate two entirely separate online feedback platforms in the future in order to find workers with a genuine intrinsic motivation.

APPENDIX

REFERENCES