

To Play or Not to Play: Interactions between Response Quality and Task Complexity in Games and Paid Crowdsourcing

Markus Krause, René Kizilcec

UC Berkeley ICSI, Stanford University
 markus@icsi.berkeley.edu, kizilcec@stanford.edu

Abstract

Digital games are a viable alternative to accomplish crowdsourcing tasks that would traditionally require paid online labor. This study compares the quality of crowdsourcing with games and paid crowdsourcing for simple and complex annotation tasks in a controlled experiment. While no difference in quality was found for the simple task, paid contributors' response quality was substantially lower than players' quality for the complex task (92% vs. 78% average accuracy). Results suggest that crowdsourcing with games provides similar and potentially even higher response quality relative to paid crowdsourcing.

Introduction

Digital games have become recognized as a viable alternative to accomplish crowdsourcing tasks that would otherwise require paid online labor (Pe-Than, Goh, & Lee, 2014). In recent work, digital games have been used to collect natural language data (Simko, Tvarozek, & Bielikova, 2011; Musat & Faltings, 2013; Thisone et al., 2012) and train machine learning systems (Barrington, Turnbull, & Lanckriet, 2012). Despite the frequent use of digital games in crowdsourcing, there has not been a formal comparison of the quality of responses from players of a gaming portal and paid contributors from a crowdsourcing platform. If responses collected from games are of superior quality than responses from paid crowdsourcing, it could offset the additional cost incurred by developing a game for crowdsourcing, rather than using a platform for paid online labor (Krause 2014).

Besides differences in the computer interface, games and paid crowdsourcing also differ in terms of the characteristics of the population of game *players* and paid *contributors*. Most notably, these two populations have different incentives to engage in the task: financial incentives for paid contributors and entertainment for players. The role of intrinsic and extrinsic motivations in crowdsourcing has



Figure 1: Game for the simple image annotation task. Player enter a keyword for the shown image in the text field on the left. The current score and a list of images for which the player gave the most common answer (right).

been the subject of prior investigation (Findley et al., 2012). Mao et al. (2013) also examined differences between voluntary and paid contributions for different tasks. In contrast to volunteering contributors who are motivated by the task and its underlying goal, game players are rarely aware of the task and they engage for their own entertainment.

The contribution of this work is to provide a first insight into qualitative differences between crowdsourced responses gathered from paid contributors on a crowdsourcing platform and players completing an equivalent task in the context of a game. In a field experiment, we investigate differences in the quality of responses from games and paid crowdsourcing for different tasks. The first task is a simple image annotation task in which participants label images with relevant tags. The second task is more complex: participants annotate web page excerpts with questions they think can be answered with the given excerpt. On this complex task, players achieve signif-

icantly higher average quality ratings ($M = 0.92$) than paid contributors ($M = 0.78$). On the simpler image annotation task, the average quality ratings of paid contributors ($M = 0.96$) are not significant different from players ($M = 0.92$).

Related Work

In an ongoing line of research comparing experts to crowds of untrained contributors, studies have investigated non-experts’ ability to predict election outcomes (Sjöberg, 2009) and which news stories would become popular (Hsieh et al., 2013). Rashtchian and Young (2010) compared the quality of captions provided by paid contributors and expert annotators. Other work examined the potential benefits of combining expert and non-expert responses to achieve results of high quality with fewer overall responses (Kandasamy et al., 2012). This line of work informs the choice between high-quality yet costly expert contributions on the one hand, and somewhat lower quality but less expensive crowd contributions on the other hand.

Human computation can be an effective alternative for more complex natural language tasks, such as for semantic similarity analysis (Batram et al. 2014), recognizing entailment, and word sense disambiguation (Snow et al., 2008). Similar comparisons between expert and non-expert contributors have been conducted in this domain. Oosten and Hoste (2011) compared expert and non-experts on judging the readability of texts, and Anastasiou and Gupta (2011) compared translations of texts by crowds to machine generated translations.

There are a number of practical, real-world applications that rely on human computation for natural language understanding. *VizWiz* (Bigham et al., 2010), for instance, allows visually impaired people to post images along with a question to an online platform and receive answers in near real-time. Similarly, Jeong and colleagues (2013) social question answering system provides crowdsourced answers on Twitter, which were found to be of similar quality as answers from peers. An example of a digital game with a natural language understanding task for collecting practical information from the crowd is *Verbosity* (von Ahn, Kedia, & Blum, 2006), in which players generated common-sense knowledge for the purpose of making computer programs more intelligent. An early example of game-based crowdsourcing is a game designed to collect words and phrases for describing images, called the *ESP* game (von Ahn & Dabbish, 2004). The properties of responses collected in these two games were not compared to ones gathered from paid contributors competing similar tasks.

While there exist numerous studies comparing experts to non-experts and human-generated to machine-generated responses, this brief review of the literature highlights the need for a comparison of the quality of responses from

paid crowdsourced tasks and tasks completed in games. We therefore pose the followign research question:

RQ1: Is the response quality higher in games or in paid crowdsourcing?

Considering the additional costs incurred by developing a game for crowdsourcing (Krause et al. 2010), responses collected from games should exhibit superior qualities to justify the expense. Games can be highly complex and human computation games have already demonstrated their potential in solving complex problems (Cooper et al., 2010). This bears the question whether games are more suitable for complex human computation problems:

RQ2: How does task complexity influence the difference in response quality between games and paid crowdsourcing?

Study Design

The experiment follows a between-subjects design with two factors: task *complexity* (simple versus complex) and *incentive* (payment versus entertainment). The *simple* task is to annotate images with labels—a canonical task in crowdsourcing. The *complex* task is to annotate web resources with a question that could be answered using the resource. The incentive for completing the task is either a *payment* in the case of paid crowdsourcing or the entertainment value derived from playing a digital game. Each participant in the study experiences one of the four conditions.

Participants

We collected data from 1075 players and 187 paid contributors. We recruited players via two major online gaming platforms with a substantial user base (*kongregate* and *newgrounds*)¹. All paid contributors were recruited through Crowdfunder. We randomly selected 50 participants (with at least 10 responses) from each condition to estimate the response quality in each condition. As we hand labeled responses to estimate quality we did not analyze all collected samples. Table 1 shows the number of participants in each experimental condition.

Task	Incentive	Participants	Analyzed
Simple	Payment	89	50
	Game	837	50
Complex	Payment	98	50
	Game	238	50

Table 1: Distribution of participants within conditions.

¹ *Kongregate* (www.newgrounds.com) has ~1.4 million daily visitors, *newgrounds* (www.newgrounds.com) 850,000 (approximations taken from *wolframalpha.com*)

Measures

The dependent variable in this study is perceived response quality. Measuring response quality in a human computation scenario is inherently challenging as it is not possible to precompute correct results, or gold data. We therefore ask two human judges to estimate the response quality on a scale from 0 to 1. On a webpage, judges see the initial request (either an image or a web site excerpt) and the response given by the paid contributor or game player. The interface is identical to the interface used by paid contributors, except that judges see an additional slider to report the perceived quality of each response. All judges were independently selected and the process was blind and randomized, i.e. judges did not know the condition of the response they rated and were presented with responses from all conditions at random. We had a total of eight judges evaluating 500 responses in each group (2,000 total). We report the interrater agreement for each group using Krippendorff's Alpha (Krippendorff 1970).

Procedure

Immediate feedback is a characteristic feature of games, which has been found to improve response quality in paid crowdsourcing (Dow et al., 2012). Moreover, in the absence of feedback in games (i.e., unsupervised scenarios), players were found to be less reliable than when feedback was provided (Von Ahn & Dabbish 2008; Prestopnik, Crowston, & Wang 2012). These findings suggest that feedback is a critical ingredient for obtaining quality responses from both game-based and paid crowdsourcing. Paid contributors also receive feedback in crowdsourcing, but typically not until the task is completed and the contributor has been compensated. This illustrates that the availability of immediate feedback is a potential confounding factor when comparing paid crowdsourcing with human computation games. For a clean comparison, we provide the same feedback available in our games to paid contributors. We redirect contributors to our website as Crowdflower does not support custom quality management.

Condition: Simple Task, Entertainment Incentive

We collected approximately 28,000 responses from 867 players in 10 days with our game. The dataset used for the experiment consisted of approx. 3,600 images. To compose this set we selected 160 nouns from a list of common words of the English language (Noll & Noll, 2006). For each noun, we retrieved a set of images via *Google* image search. Each image was labeled with the most frequently occurring noun on the image source website. We added each image with its label to a database to bootstrap our



Figure 2: The feedback screen shows a calculated score (number in blue), how many other players also responded with the same label (percentage in blue), and the top answer for this image (bottom).

quality management system. The image-labeling game was similar to the *ESP* game (von Ahn & Dabbish, 2004) with the general game idea inspired by the successful *Family Feud* game show. In this show, two groups compete against each other. The goal of the game is to find the most popular responses to a survey question that was posted to a group of 100 individuals. Each request only takes a few seconds to complete ($M = 0.08$, $SD = 0.03$). When a player responded with the most frequent answer to the shown image, we added the image to a list. The last entry of this list was shown to the player on the right side of the game screen. Figure 1 shows a screenshot of the game interface.

The game showed statistics to players during the game: the number of rounds played, the number of times the player found the top answer, and a score for the last response. The game also showed a feedback screen for each response as seen in Figure 2. The game responded to the use of swear or slang words, and repetition of the same label. It also showed spelling errors and auto corrections if available. To report a score for the entered label, the most similar label for the current image was provided as additional feedback. Labels with strong semantic similarity were treated as equal.

Condition: Simple Task, Payment Incentive

We published the same image-labeling task on *Crowdflower* paying \$0.01 for each image label (chosen to provide an ethical average wage of \$7/hour), allowing contributors to label up to 100 images in batches of 10. We collected 3,756 responses from 89 paid contributors. To provide the same level of immediate feedback as in our game condition, we asked contributors to follow a link to our website to complete the task. The website showed a batch of 10

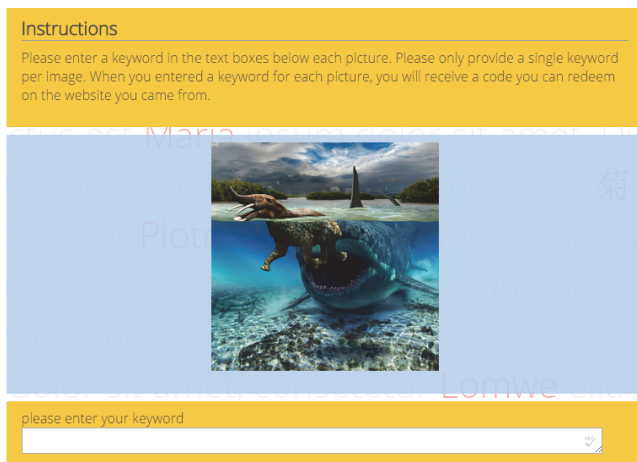


Figure 3: The image task as seen by a paid contributor. A short section (top) gives basic instructions. Contributor enters a single keyword or term in the text field.

images. The website used the same tool to generate feedback as our game. We reported a quality estimate between 1 and 100 for each response and gave notifications if inappropriate terms such as swear and slang words were used. The feedback appeared immediately after a contributor submitted a label.

Some feedback was an inherent part of the game mechanics such as terms entered by other players. We did not report this information in the payed condition. A screenshot of the interface can be seen in Figure 3. After a contributor entered a label for each image, they received a code to redeem on Crowdfunder to receive their payment. Contributors received this code only if their average predicted quality was higher than 50. We used this measure to protect the task against low quality contributions. We allowed contributors to revise their responses in these cases.

Condition: Complex Task, Entertainment Incentive

We collected approx. 2,500 responses from 238 players in 10 days. The task was to generate questions based on text excerpts, such that the answer to the generated question is contained in the excerpt (Aras et al. 2010). The game developed for crowdsourcing this task was inspired by the successful 1960s quiz show *Jeopardy*. In this show, players received clues in the form of answers and were asked to phrase their response as a question. Our game gives a short introduction to the game (Figure 4) and asks players to select a category and a level of difficulty from a table (Figure 5). The difficulty levels were computed based on implicit user feedback, such as the response's length and complexity, and the time taken to enter the response. For the initial ranking we calculate the *Flesch Reading Ease* (Kincaid et al., 1975). Despite this freedom to choose in which order to complete the tasks, players had to complete tasks of all categories and all difficulties to win the game. In the paid

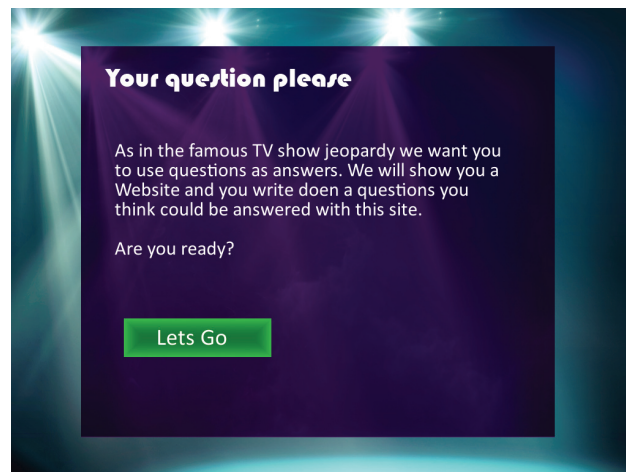


Figure 4: Instruction screen of the game for the complex annotation task.

version, the order in which workers completed tasks was pre-determined. We expect this difference in ordering to be inconsequential, as the tasks themselves remain constant between conditions.

The text excerpts presented to players were article summaries from the Google news aggregator. After choosing a category and difficulty, the game showed players a corresponding news summary with an input field to enter their response. Players had 30 seconds to complete each task (Figure 6). If the response was considered a valid question, the player earned points according to the level of difficulty. We assessed the response quality based on the number of unique terms, spelling errors in the response, and its grammatical structure (Kilian et al. 2012). The system returns quality estimates in three categories low (0) mediocre

	Entertainment	Politics	Science	Sports
100	100	100	100	100
200	200	200	200	200
300	300	300	300	300
400	400	400	400	400

Figure 5: Task selection screen of the game for the complex annotation task. The first row gives the categories a player can choose from. The numbers indicate difficulty and the amount of points rewarded for a correct answer.

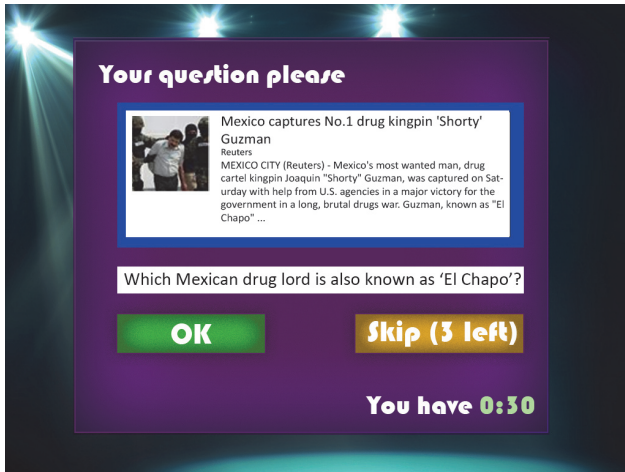


Figure 6: Player enter a question they think can be answered with the shown news snippet. The counter on the lower right shows the remaining time. Player can also skip a snippet with the yellow button.

(1) high (2). This feedback method is generic for tasks that require contributors to enter text in English and does not require ground truth data. The quality prediction used to provide feedback highly correlates with the mean of our human judges (Spearman’s $Rho > 0.7$). In comparisons, the human judges achieved a correlation of around 0.8 among themselves, which highlights the quality of the automated feedback mechanism. Responses were ranked by asking players to select the most suitable of three player-generated responses for a given news summary (Figure 7). The player could also skip in case all responses were unacceptable. This related task was randomly shown to players as a bonus level when selecting a table cell.

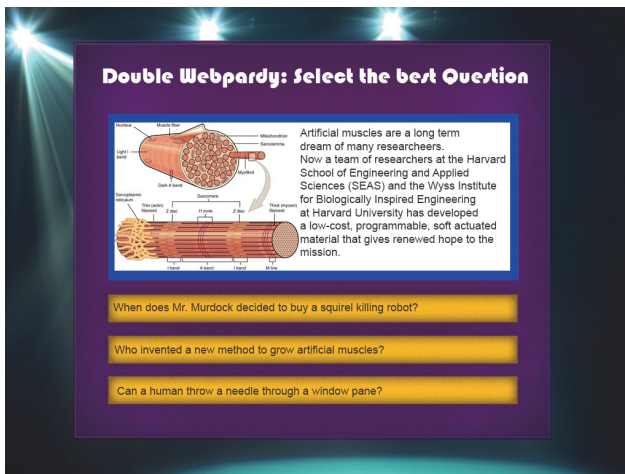


Figure 7: Double Jeopardy screen. The player has a certain amount of time to select as many appropriate questions for the shown Fragments as possible.



Figure 8: Crowdsourcing interface for the more complex web-fragment annotation task.

Condition: Complex Task, Payment Incentive

For the paid crowdsourcing condition, we recruited 98 contributors via Crowdfunder and collected 1,286 responses. We paid \$0.15 per annotation (chosen to provide an ethical average wage of \$7/hour) and allowed contributors to annotate up to 50 web page excerpts in batches of five. To provide the same feedback as in the game, we redirected contributors to our website. The web-interface was almost identical to the simple annotation task shown in Figure 8, but showed five web page excerpts instead of the ten images. We asked contributors to enter a question that could be answered with the given web page excerpt and applied the same quality management method as in our game. Paid contributors saw a message with a quality estimate immediately after submitting each response.

In contrast to the game contributors did not have a time limit to enter their response. We again protect our task against low quality responses. Contributors received their code only if their average predicted quality was higher than 1 and allowed contributors to revise their responses.

Results

To analyze the response quality we randomly selected 50 participants from each group and randomly selected 10 responses from each participant. At least two out of 8 judges rated each response on a scale from 0 to 1 as explained in the measurement section. The agreement between judges was high for the complex task (Krippendorff’s $\alpha = 0.78$) and even higher for the simple task ($\alpha = 0.85$). These values indicate substantial agreement (Krippendorff 1970).

Figure 9 shows the average response quality in each condition with bootstrapped 95% confidence intervals. We use an Analysis of Variance to investigate main and interaction effects (see Table 2). In accordance with Harwell et al. (1992) and Schmider et al. (2010) we assume our group

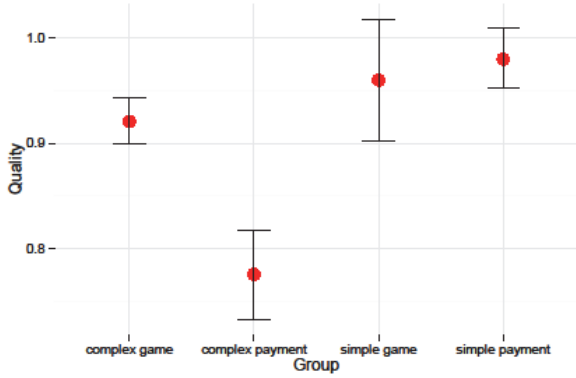


Figure 9: The 95% confidence intervals of the means for each of the four groups. The intervals are calculated from 100,000 bootstrap samples (Efron and Tibshirani 1986).

sample size ($N=50$) and our substantial effect sizes (Cohen's $d > 0.6$ on average) to be sufficient to meet ANOVA's normality criterion.

We conducted a Levene's test for homogeneity and did not find a significant deviation of our data from the equal variance assumption $F(3,196) = 1.87, p = 0.13$. Consequently, we use a series of Welch two sample t -tests as our post-hoc tests and apply Holm's method (Holm 1979) to account for multiple comparisons. The Welch test uses an estimate of degrees of freedom (Df) that can be much lower than the actual sample size. We report Df in integer precision.

	Df	SS	MS	F	p	$sig.$
(I)ncentive	1	0.22	0.22	8.45	0.003	**
(T)ask	1	1.53	1.53	60.09	0.000	***
IxT	1	0.27	0.27	10.70	0.001	**
Residuals	197	4.01	0.03			

Table 2: ANOVA results of main and interaction effects between the factors incentive structure (game and payment) and task complexity (web-fragment and image annotation).

Higher Complexity lower Response Quality

As seen in Table 2 we found that the presumed complexity difference of both tasks had a significant impact on the response quality. This seem to be obvious but it illustrates that our initial assumption that our tasks differ in complexity is in fact true. This finding is in-line with results from a survey on Crowdfunder. Upon completion of a task, Crowdfunder asks contributors to take a survey. One question in this survey regards the *ease of job* rated on a scale from 0 (hard) to 5 (easy). The less complex image annotation task received an average score of 4.5 ($N=43$) the more complex web-annotation task a score of 3.3 ($N=51$).

Players have a higher Response Quality

The incentive structure also has a significant impact on response quality as seen in Table 2 and Figure 10. Players (M

$= 0.93, SD = 0.10$) have a significantly $T(111) = 3.16, p = 0.008, d = 0.44$ higher average response quality than paid contributors ($M = 0.86, SD = 0.21$). For the image-annotation task, the difference in means between paid contributors ($M = 0.98, SD = 0.12$) and players ($M = 0.96, SD = 0.12$) is not significant $T(85) 0.42, p < 0.68, d = 0.534$. For the complex annotation task on the other hand players have a significantly higher response quality ($M = 0.92, SD = 0.09$) than paid contributors ($M = 0.78, SD = 0.21$) $T(52) = 1.21, p < 0.001, d = 0.534$. This is an increase of almost 18% in response quality.

Discussion

This study sheds light on the question of how the quality of crowdsourcing with games compares to paid crowdsourcing. Based on our experiment with two different tasks of varying complexity and controlled populations, we found games to generate higher quality data for complex tasks. The quality increase in our case was almost 18% for the complex web-fragment annotation task. We did not find such a significant difference in the average response quality for the less complex image-annotation task.

Given these results, we can answer our initial research question **RQ1**. For our complex task player have a significantly higher response quality. It is also possible to respond to **RQ2** as there is a significant interaction between task complexity and incentive structure.

A possible explanation of this interaction is that players are more selective than paid contributors are. Player chose our games for their entertainment if the game and the underlying task does not appeal to them they will not chose to play the game or quit to play soon. In contrast, paid contributors are more interested in payment. As long as the job pays the bills, it is not as important if you like it.

An indication for a higher selectiveness in players is the number of players for both games. With 837 players in ten

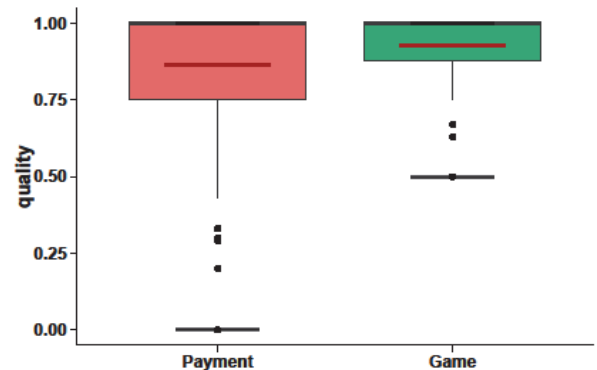


Figure 10: Comparison between the incentives game and payment in terms of response quality with both annotation tasks combined.



	Complex Game	Complex Task	Simple Game	Simple Task
		Mexico captures No.1 drug kingpin 'Shorty' Guzman Reuters MEXICO CITY (Reuters) - Mexico's most wanted man, drug cartel kingpin Joaquin "Shorty" Guzman, was captured on Saturday with help from U.S. agencies in a major victory for the government in a long, brutal drugs war. Guzman, known as "El Chapo" ...		
High	Who was captured by the Mexican police and on their number one most wanted list?	Who was responsible for the arrest of Joaquin Guzman in Mexico?	Megalodon	Mammoth
Low	My nickname was "El Chapo"	How did he even get here?	My gold fish eating the cat	word

Table 3: Examples of responses. The first row indicates the condition. The second row contains a depiction of the actual request. In the third row the table shows high rated answers and the fourth low rated answers.

days, the simple image annotation game had almost four times more players than the more complex game with only 238 players. One could attribute this difference to a generally higher quality of the first game, yet both games received similar ratings on both portals. This seems to underline that players deliberately chose a game. This selectiveness is especially important for complex tasks that require more commitment than simple ones. This finding would also be in line with other studies that successfully use games as a vehicle for complex tasks (Cooper et al. 2010).

Future Work

In our experiment, we measured perceived quality with human judges. An equally interesting question is how exhaustive both methods can explore the possible answer space of a task. We will investigate coverage in a future paper in detail. We assume that players are more likely to explore a wider range of possible responses. In a gaming situation humans more tend to explore a behavior less prevalent in paid work. Measuring coverage is easier with tasks for which all or at least most possible responses are known in advance. Factual tasks such as *Name all presidents of the United States* lend themselves more easily to investigate coverage. On the other hand, such questions might be too restrictive to explore the full potential of a method.

Another factor that would be relevant to explore is expertise. As mentioned in the related work section many studies already compare experts and contributors it would therefore be a natural addition to see how well players perform compared to experts. Furthermore, in the presented experiment we provided immediate feedback to all participants. As argued in the related work section immediate

feedback has a positive effect on response quality. It is an interesting question if there is a significant interaction between the two investigated factors (task complexity and incentive structure) and the level of provided immediate feedback.

References

- Anastasiou, D., and R. Gupta. 2011. "Comparison of Crowdsourcing Translation with Machine Translation." *Journal of Information Science* 37 (6): 637–59.
- Aras, Hidir, Markus Krause, Andreas Haller, and Rainer Malaka. 2010. "Webpardy: Harvesting QA by HC." In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*, 49–53. New York, New York, USA: ACM Press.
- Barrington, Luke, Douglas Turnbull, and Gert Lanckriet. 2012. "Game-Powered Machine Learning." *Proceedings of the National Academy of Sciences of the United States of America* 109 (17): 6411–16.
- Batram, Nils, Markus Krause, and Paul-olivier Dehaye. 2014. "Comparing Human and Algorithm Performance on Estimating Word-Based Semantic Similarity." In *SoHuman'14 Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation*, 131–39. Barcelona, Spain: Springer Berlin Heidelberg.
- Bigham, JP, Chandrika Jayant, Hanjie Ji, and Greg Little. 2010. "Vizwiz: Nearly Real-Time Answers to Visual Questions." In *UIST '10 Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, 333–42. ACM Press.
- Cooper, Seth, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, and Zoran Popović. 2010. "Predicting Protein Structures with a Multiplayer Online Game." *Nature* 466 (7307): 756–60.
- Dow, Steven, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. "Shepherding the Crowd Yields Better Work." *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12*. New York, New York, USA: ACM Press, 1013.

- Efron, Bradley, and Robert Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 2 (1): 54–75.
- Findley, MG, MC Gleave, RN Morello, and DL Nielson. 2012. "Extrinsic, Intrinsic, and Social Incentives for Crowdsourcing Development Information in Uganda: A Field Experiment." In *WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference*, 326–35. Provo, UT, USA: ACM Press.
- Harwell, M. R., E. N. Rubinstein, W. S. Hayes, and C. C. Olds. 1992. "Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases." *Journal of Educational and Behavioral Statistics*.
- Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.
- Hsieh, Chu-cheng, Christopher Moghbel, Jianhong Fang, and Junghoo Cho. 2013. "Experts vs the Crowd: Examining Popular News Prediction Performance on Twitter." In *Proceedings of ACM KDD Conference*. Chicago, USA: ACM Press.
- Jeong, JW, MR Morris, Jaime Teevan, and Dan Liebling. 2013. "A Crowd-Powered Socially Embedded Search Engine." In *Proceedings of ICWSM 2013*, 263–72.
- Kandasamy, DM, K Curtis, Armando Fox, and David Patterson. 2012. "Diversity within the Crowd." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12*, 115–18. Seattle, Washington, USA: ACM Press.
- Kilian, Niklas, Markus Krause, Nina Runge, and Jan Smeddinck. 2012. "Predicting Crowd-Based Translation Quality with Language-Independent Feature Vectors." In *HComp'12 Proceedings of the AAAI Workshop on Human Computation*, 114–15. Toronto, ON, Canada: AAAI Press.
- Kincaid, J Peter, Robert P Fishburne, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Air Station Memphis-Millington, TN, USA: National Technical Information Service, Springfield, Virginia.
- Krause, Markus. 2014. *Homo Ludens in the Loop: Playful Human Computation Systems*. Hamburg, Germany: tredition GmbH, Hamburg.
- Krause, Markus, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. "Frontiers of a Paradigm." In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*, 22–25. New York, New York, USA: ACM Press.
- Krippendorff, Klaus. 1970. "Estimating the Reliability, Systematic Error and Random Error of Interval Data." *Educational and Psychological Measurement* 30 (61): 61–70.
- Mao, Andrew, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. 2013. "Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing." In *HComp'13 Proceedings of the AAAI Conference on Human Computation*, 94–102. Palm Springs, CA, USA: AAAI Press.
- Musat, CC, and Boi Faltings. 2013. "A Novel Human Computation Game for Critique Aggregation." In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 86–88. AAAI Press.
- Noll, Paul, and Bernice Noll. 2006. "3000 Most Commonly Used Words in the English Language (USA)." *"Clear English" a Book on Teaching and Learning English*. Pleasant Hill, OR, USA. <http://www.paulnoll.com/Books/Clear-English/English-3000-common-words.html>.
- Oosten, Philip Van, and V Hoste. 2011. "Readability Annotation: Replacing the Expert by the Crowd." In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 120–29. Association for Computational Linguistics.
- Pe-Than, Ei Pa Pa, Dion Hoe-Lian Goh, and Chei Sian Lee. 2014. "A Typology of Human Computation Games: An Analysis and a Review of Current Games." *Behaviour & Information Technology* 00 (00). Taylor & Francis: 1–16..
- Prestopnik, Nathan, Kevin Crowston, and Jun Wang. 2012. "Exploring Data Quality in Games With a Purpose Theory: Gamification and Games With a Purpose." In *Proceedings of the 2012 iConference*, 1–15. ACM Press.
- Rashtchian, C, and P Young. 2010. "Collecting Image Annotations Using Amazon's Mechanical Turk." In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139–47. Stroudsburg, PA, USA.
- Schmider, Emanuel, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. "Is It Really Robust?: Reinvestigating the Robustness of ANOVA against Violations of the Normal Distribution Assumption." *Methodology* 6 (4): 147–51.
- Simko, J, M Tvarozek, and M Bielikova. 2011. "Little Search Game: Term Network Acquisition via a Human Computation Game." In *Proceedings of the 22nd International Conference on Hypertext HT'11*, 57–61. Eindhoven, The Netherlands: ACM Press.
- Sjöberg, L. 2009. "Are All Crowds Equally Wise? A Comparison of Political Election Forecasts by Experts and the Public." *Journal of Forecasting*.
- Snow, Rion, Brendan O Connor, Daniel Jurafsky, Andrew Y Ng, Dolores Labs, and Capp St. 2008. "Cheap and Fast — But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–63. Association for Computational Linguistics.
- Thisone, Musat, Claudiu-Cristian, Alireza Ghasemi, and Boi Faltings. 2012. "Sentiment Analysis Using a Novel Human Computation Game." In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and Their Applications to NLP*, 1–9. Association for Computational Linguistics.
- Von Ahn, Luis, and Laura Dabbish. 2004. "Labeling Images with a Computer Game." In *CHI '04 Proceedings of the 22nd International Conference on Human Factors in Computing Systems*, 319–26. Vienna, Austria: ACM Press.
- Von Ahn, Luis, and Laura Dabbish. 2008. "Designing Games with a Purpose." *Communications of the ACM* 51 (8). ACM: 58–67.
- Von Ahn, Luis, Mihir Kedia, and Manuel Blum. 2006. "Verbosity: A Game for Collecting Common-Sense Facts." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 22:78. Montreal, Quebec, Canada: ACM Press.