Worker Perception of Quality Assurance Mechanisms in Crowdsourcing and Human Computation Markets

Completed Research Paper

Thimo Schulze
University of Mannheim
Business School
Germany
schulze@wifo.uni-mannheim.de

Dennis Nordheimer
University of Mannheim
Business School
Germany
dennis.nordheimer@wifo.uni-mannheim.de

Martin Schader
University of Mannheim
Business School
Germany
martin.schader@uni-mannheim.de

ABSTRACT

Many human computation systems utilize crowdsourcing marketplaces to recruit workers. Because of the open nature of these marketplaces, requesters need to use appropriate quality assurance mechanisms to guarantee high quality results. Previous research has mostly focused on the statistical aspects of quality assurance. Instead, we analyze the worker perception of five quality assurance mechanisms (Qualification Test, Qualification Restriction, Gold Standard, Majority Vote, Validating Review) according to subjective (fairness, offense, benefit) and objective (necessity, accuracy, cost) criteria. Based on theory from related areas like labor psychology, we develop a conceptual model and test it with a survey on Mechanical Turk. Our results show big differences in perception, especially with respect to Majority Vote which is rated low by workers. On the basis of these results, we show implications for theory and give requesters on crowdsourcing markets the advice to integrate the worker view when selecting an appropriate quality assurance mechanism.

Keywords

Quality Assurance, Survey, Crowdsourcing, Human Computation, Mechanical Turk, Majority Vote, Labor Psychology

INTRODUCTION

Human Computation has emerged as a powerful new paradigm where humans and computers work together to solve hard problems that neither of them can solve alone (Quinn and Bederson, 2011). For example, many tasks like content development, data tagging, natural language processing, image understanding, or knowledge representation are still very hard or impossible to solve by computers, yet very easy to perform by humans (Yampolskiy, 2012). The human computation paradigm attempts to use the combined capabilities of humans and computers to solve these problems. Many human computation systems use crowdsourcing marketplaces to recruit workers.

The term "Crowdsourcing," the act of outsourcing tasks to a large and undefined group of people in the form of an open call (Howe, 2008), is used for a wide variety of applications in areas like collaborative knowledge creation (Wikipedia), open innovation (Chesbrough, 2003), or competition markets (Leimeister et al., 2009). For this paper, however, we focus on a narrower definition of crowdsourcing marketplaces that can be used to dynamically recruit workers for human computation systems. The most prominent example is Amazon Mechanical Turk (MTurk), a marketplace for work where businesses and developers ("requesters") can get access to an on-demand, scalable workforce ("workers"). Requesters post Human Intelligence Tasks ("HITs") that are then self-selected by the workers who perform the work and submit the results back to the system. Workers can flexibly select work of different kinds and from different requesters and are paid on a per-task basis where most tasks are micro tasks that can be completed in a few minutes.

Because of the novel open call format where anonymous workers self-select the tasks, requesters have to overcome the challenge of ensuring high quality results. Some workers submit incorrect results because they are not qualified for the work, make careless mistakes, or are simply trying to submit random results to game the systems (Downs et al., 2010). Several statistical quality assurance mechanisms (QAMs) have been proposed to manage the quality in this setting. First, the worker population that is allowed to work on the task can be restricted by qualification restrictions or qualification tests (Kazai, 2010). For tasks like data categorization and labeling, which have a limited answer space and one objectively verifiable correct solution, majority vote approaches can be applied where the same task is performed by multiple workers and the answers are compared (Kern et al., 2011). Some providers like CrowdFlower use "gold standard" units where tasks with known answers are injected into the work stream to infer the quality of a worker on other tasks (Oleson et al., 2011). In the case where different answers can be correct (like content creation or translation), peer review approaches can be employed where the submitted results are verified by different workers (Sun et al., 2011).

Most of these approaches treat the human submissions to the tasks as mere noisy input data that has to be cleaned up with statistical techniques in order to reduce costs giving desired quality levels or to maximize quality giving budget restrictions. However, little research has been done on the impact that these systems have on the workers that perform the tasks. Given a specific quality assurance approach, good workers might be discouraged to put full effort into performing a task or even decide to not perform the task at all; thus the initial data quality that can be utilized in the statistical approaches is reduced. In this case, requesters might be inclined to use different approaches that are perceived more positively by workers. As a first step in exploring this area, we want to answer the following research question:

- How do workers on crowdsourcing marketplaces perceive different quality assurance mechanisms according to subjective and objective criteria?

To answer this question, we develop a preliminary conceptual model based on related areas and test it in a quantitative survey on MTurk.

RELATED WORK AND THEORETICAL FOUNDATIONS

Comprehensive work has been performed on the theoretical and practical foundations of quality assurance in crowdsourcing. But to the best of our knowledge, the perception of quality assurance mechanisms has not yet been studied in the context of crowdsourcing markets. However, similar approaches have been analyzed in other areas like business, economics, or psychology and can therefore support the theoretical foundation of our model.

Quality Assurance Mechanisms in Crowdsourcing

Qualification Test (QT). Many crowdsourcing platforms offer requesters the ability to design qualification tests. Workers need to pass these tests before they are allowed to work on actual tasks. This mechanism can be used to train workers, illustrate what answers are expected, or to test the skills and abilities of workers. While these qualification tests can filter out bad workers, it could also discourage honest workers who are not willing to invest time in a test with unknown prospects of future earning potential (Wais et al., 2010). On the other hand, workers could put great effort into passing the test and then reduce this effort on the actual task.

Qualification Restriction (QR). Qualification-based restrictions limit the approved worker pool by certain qualitative or quantitative metrics (Schulze, Krug, et al., 2012). These are usually combinations of the quantity of historical work and the quality of or satisfaction with historical work; e.g., MTurk uses the "HIT approval rate" and the number of "total approved HITs." Workers can usually acquire these reputations by completing other tasks on the platform. One problem with global restrictions lies in the fact that the skills of workers might be task specific, e.g., workers who are experienced in image tagging are not necessarily skilled in translation as well. Also, worker quality might change over time.

Gold Standard (GS). Currently, many crowdsourcing providers typically use the gold standard data sets to evaluate worker quality. The main idea behind this method is that tasks with known answers are mixed into the stream of regular tasks that workers process. If a worker's responses deviate significantly from the gold standard, this procedure can be employed to automatically reject such poorly performed contributions or to help the worker to learn what is required. Obviously, this method belongs to test mechanisms that can be used throughout completing the tasks to filter out non-serious workers. Thus, the gold standard questions should be selected carefully so that they cannot be differentiated from regular tasks and do not dupe good workers.

Majority Vote (MV). A substantial body of the literature focuses on quality assurance for micro tasks, such as data categorization or image labeling. A widespread method to evaluate results of these tasks is majority voting (Sorokin and Forsyth, 2008). This technique compares or aggregates multiple results provided by different workers to the same task in order to derive a single correct result or the result with highest probability of correctness (Ipeirotis et al., 2010; Jung and Lease, 2011). One drawback of this approach lies in the fact that, if worker are not trusted or the result is not of desired correctness, the system passes the same task to other workers and causes more costs. Another risk is that the consolidation of two or more false answers can vote down the correct result.

Validating Review (VR). In case that the tasks are not deterministic (e.g., language translation or creative design) so that multiple results can be considered valid, validating review (also known as peer review) is used for quality assurance. This method leverages a reviewer or a group of reviewers to assess the results submitted by other workers (Kern et al., 2010). Validating review can discourage good workers since correct results can be voted down by incorrect reviews.

To keep our selection simple, we only use these five mechanisms for our study. There are other ways, e.g., iterative approaches where a second worker improves the results of the initial submission (Little et al., 2010), improved peer review, where majority vote is used to assess the work quality instead of just a single vote (Kern et al., 2010), or techniques that are transparent to the user like using implicit behavioral measures to predict task performance (Rzeszotarski and Kittur, 2011).

Perception of Quality Assurance Mechanisms

Cornell and Welch (1996) establish a simple model that explains how discrimination can develop during pre-screening of people and how the similarity of the general background between applicants and employers influences hiring decisions. The goal behind the screening process is to select the best workers from the pool of job applicants by disclosure of their intangible characteristics. The screening process should be performed primarily in sectors where inferred worker quality is of high importance and where it is more efficient to pre-screen employees than to measure their performance afterwards (Cornell and Welch, 1996).

Harris et al. (1990) analyze the perception of different pre-employment screening procedures by human resource practitioners. Respondents were requested to rate pre-employment tests on attributes like offensiveness, discrimination, accuracy, fakeability, costs, etc. for different job types. Targeted interviews and accomplishment tests were perceived to be most accurate and least expensive. None of the applicant screening procedures were viewed as particularly discriminatory whereas biological and physiological measures were rated as offensive, least accurate, and most costly. Honesty and personality tests were seen as relatively easy to fake whereas cognitive ability tests were regarded to be hardly fakeable (Harris et al., 1990).

In contrast to the above study, Ryan and Sackett (1987) analyze the perception of screening tests from the applicants' perspective. Their study addresses the reaction of applicants to honesty tests and examines the impact of taking such a test on employees' perception of the company image. They ask students about fairness of honesty tests, feelings of being offended, privacy issues, rejection of job offers, etc. Most participants consider the application of honesty tests as appropriate and not detrimental to the company image (Ryan and Sackett, 1987).

For a comprehensive review of the research on applicant perceptions of screening processes we refer to the work of Ryan and Ployhart (2000). Among other things, they summarize what are perceptions having been studied, what are parameters of perceptions, and what are the outcomes of holding more positive or negative perceptions.

PRELIMINARY RESEARCH MODEL

Since the term perception is always connected with emotions, we first examine related works in affective science, the study of emotion or affect. The primary challenge is to distinguish between different layers of emotions and, consequently, between the ones of worker perception of QAMs. Calhoun (2004) points out that there are two general types of emotions: subjective and objective emotions. While subjective emotions are defined to be biographical charged, objective emotions are described to be epistemic. In contrast to subjective emotions, epistemic objective emotions project the truth and reality. However, pure objectivity does not exist since everything is processed through the individual point of view. Transferring statements of Calhoun (2004) to our model, we split worker perception of QAMs into the following two layers: *biographical subjective perception* and *epistemically objective perception*.

In our model, biographical subjective perception defines a layer of all subjective emotions of a worker. Each worker can develop individual emotions for the same task and thus can perceive a QAM differently. In contrast, epistemically objective

perception describes the point of view of a person without any subjective estimation. Even if it is not possible to achieve pure objectivity, it is feasible to eliminate the personal bias and incorporate the interests of a group of people (Calhoun, 2004). Accordingly, working on tasks in crowdsourcing platforms workers cannot act absolutely objective but they can judge consequences of a task's execution from a general point of view. [1]

The next step in the development of our model deals with the identification of specific attributes of the subjective and objective perception. As a starting point for this, we use the work of Lomas (1994) which establishes several characteristics of subjective and objective perception: Subjective perception can be characterized as personal, internal, emotional, intuitive, limitless, etc., while objective perception is defined to be impersonal, external, intellectual, scientific, limited, etc. The work of Lomas (1994) was taken as a basis for our model and extended by other properties established in the related literature in the previous section in the context of the QAMs considered. Figure 1 shows the resulting model in which subjective perception is divided into fairness, offense, and benefit, and objective perception consists of the characteristics necessity, accuracy, and costs. The detailed classification of all characteristics with respect to each layer is given in the next two subsections.

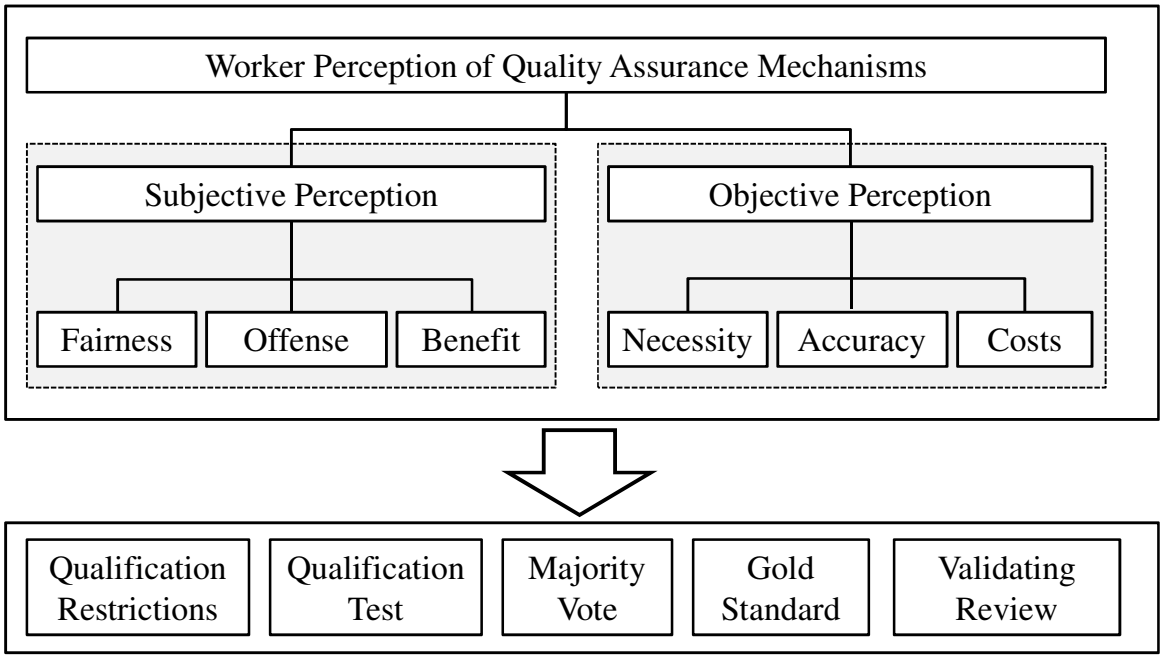

Figure 1. Preliminary Model on Worker Perception of Quality Assurance Mechanisms

Subjective Perception

Based on the findings from the previous section, as well as on further studies concerned with applicants' perception issues (Gilliland, 1993; Macan et al., 1994; Ryan and Sackett, 1987), several factors characterizing the subjective perception of different selection and screening processes can be identified. Fairness, offense, and benefit are the three core factors which emerge as most determinative and, according to Lomas (1994), can be classified into the layer of the biographical subjective perception:

[1] A very helpful reviewer points out that the distinction between subjective perception (i.e., the worker's view) and objective perception (i.e., the requester's view) could also be framed as a principal-agent problem (Holmstrom and Milgrom, 1991). Because the interests of the requester are usually not in line with the interests of the worker, an asymmetric information distribution and moral hazard might lead to several problems that are worth studying from an economic theoretical background. We see this as a very interesting direction for future research.

- *Fairness* with its various forms is the most frequently used characteristic of perception. In the research review by Ryan and Ployhart (2000) on applicants' perceptions of screening processes, this factor is considered in 21 of 40 analyzed studies. The intuitive nature of fairness is confirmed by Tabibnia et al. (2008). Since fairness in this context satisfies the essential properties of subjective perception pointed out by Lomas (1994), we group it into the layer of biographical subjective perception.
- *Offense* appears to be the next important factor related to subjective emotions perceived during a selection process or a pre-screening test. Harris et al. (1990) indicate differences between test types in terms of perceived offense and determine it as a factor seeming to affect tests' usage. The study by Ryan and Sackett (1987) shows that applicants consider certain types of pre-employment tests to be more offensive than the other ones. Aquino et al. (2006) observe that when employers follow certain rules in response to an offense feeling, employees see that fairness is implemented and feel appreciated by management. Thus, offense is closely linked to the workers' perception of fairness.
- One further factor that has the potential to influence employee performance, motivation, attitude, or satisfaction is the perception of the *benefits* offered by the management (Balkin and Griffeth, 1993; Iles et al., 1990). *Benefits* are not always measured as monetary compensations but can also represent intangible issues (Weathington and Tetrick, 2000).

Objective Perception

Objective perception represents the second layer of our model and considers the worker perception of QAMs from the point of view of the requester. Based on objective criteria for worker perception of different selection procedures pointed out in the previous section and taking into consideration the key challenge in crowdsourcing, namely to ensure high-quality processing of tasks, we recognized the following core factors in this context: *necessity*, *accuracy,* and *costs*:

- Since one of the great challenges of crowdsourcing is to extract high quality results from an a priori unknown population of workers, the necessity of some sort of QAM is obvious from the perspective of a requester (Kern et al., 2009). We include this aspect of *necessity* in the model in order to evaluate whether an individual worker also understands the need of these approaches.
- Harris et al. indicate substantial differences between the results of pre-employment screening procedures in terms of perceived *accuracy* (Harris et al., 1990). Studying this fact, Sacket and Wilk found out that performance on the test is closely linked to performance on the work (Sackett and Wilk, 1994).
- *Cost* reduction is one aspect often mentioned as a big advantage of crowdsourcing (Schenk and Guittard, 2011). However, this objective is opposed to the goal of most workers who want to earn as much money as possible (Kaufmann et al., 2011). Still, by using a certain QAM, the requester might be able to reduce the monetary amount paid to dishonest workers so that good work can be rewarded.

Tables 1 and 2 summarize the theoretical foundations by developing a definition in the context of crowdsourcing and human computation marketplaces. The definitions are partly adapted from the literature mentioned before.

In additional questions, general data about the registration period and the weekly time effort as well as the workers' HIT statistic (total approved HITs, number of assigned qualifications, HIT approval rate) and demographics (gender, age, level of education) was collected. Due to space limitations, answers for some questions are not reported in this paper. In order to filter out non-serious workers who do not really read every task, we injected different gold standard questions into our questionnaire. The results of respondents who answered at least one of the spamming questions wrongly were not used for the analysis.

The survey was published on the MTurk platform in May 2012 with the title "Survey about Quality Assurance on Amazon Mechanical Turk" with payment of $0.26 for every completed task and about 15 minutes scheduled for its completion. These values correspond to 70% of the average wage per hour of a worker on MTurk (Ross et al., 2010). In order to reduce the effects of cultural differences (Kaufmann et al., 2011), only participants form the USA were allowed. The survey was available on the platform for one week and led to 170 submitted HITs. After filtering out potential spammers, 159 responses to the full survey are used for the data analysis.

	Definition (in the Crowdsourcing context)	**Survey Item** *From the point of view of a worker, I think that this quality assurance mechanism …*
Fairness	Refers to the condition that a worker that behaves correctly is treated acceptably by the mechanism; and non-serious workers cannot take advantage of correctly behaving persons.	*... treats everybody equally, so nobody has an advantage.*
Offense	Refers to an emotional state of annoyance or anger towards the system or the requester.	*... annoys me and makes me angry.*
Benefit	Refers to the individual profit of the worker measured in a monetary or non-monetary value.	*... enables me to earn more money.*

Table 1. Characteristics of the Subjective Perception

	Definition (in the Crowdsourcing context)	**Survey Item** *From the point of view of a requester, I believe that this quality assurance mechanism is a way to ...*
Necessity	Refers to the objective need to use an approach in order to achieve a desired purpose or goal.	*... achieve overall results of higher quality.*
Accuracy	Refers to the degree of quality of the crowd output compared to the correct or optimal result.	*... motivate workers to submit better work.*
Costs	Refers to the overall monetary amount a requester has to reward to workers in order to complete a project with required quality.	*... reduce the total amount of money needed to complete a project.*

Table 2. Characteristics of the Objective Perception

DATA ANALSIS AND DISCUSSION

Data Analysis Method

We use IBM SPSS Statistics (v20) to analyze the data. First, we look at the perception of the five different QAMs by ranking them individually according to the six different aspects. We use the "Friedman" test for the ranking and pairwise "Wilcoxon signed-rank" tests to check if differences in the median are significant. We also report the median and quartiles for each test. Next, we analyze each QAM individually. Since absolute values have limited explanatory power for Likert scale data, we use these scales to compare differences based on the demographic questions (by forming two or more sub-groups for the comparison). As we only use one item per model construct, the Likert scales have to be considered as being ordinal and we employ the non-parametric Mann-Whitney-Wilcoxon-U-Test (Boone and Boone, 2012; Schulze, Indulska, et al., 2012).

Overall results

The overall descriptive results are depicted in Tables 3 and 4. Table 3 shows the quartiles of the survey results for each question that we asked for the five different QAMs. Additionally, we include the proportion of people who (strongly) agree with the statement. To clarify the results, we also order the results and display them in a different form in Table 4. Overall, the results suggest that Qualification Test, Qualification Restriction, and Gold Standard have a high standing with the workers, while Validating Review is seen neutrally. Majority Vote has the lowest assessment with the workers in most categories; it is especially perceived as being unfair and offensive and many workers do not see the benefit for the requesters at all. However, while many workers do not like Majority Vote, they realize its potential to decrease the costs for requesters.

		Qualification Test				Qualification Restriction				Majority Vote				Gold Standard				Validating Review			
		Quartiles			Agree	Quartiles			Agree	Quartiles			Agree	Quartiles			Agree	Quartiles			Agree
Worker View	Fairness	3	4	5	**72%**	3	4	4	**64%**	2	3	4	**38%**	3	4	4	**64%**	3	3	4	**64%**
	Offense	1	2	3	**13%**	2	3	4	**29%**	2	3	4	**40%**	2	2	3	**13%**	2	3	4	**13%**
	Benefit	3	4	4	**60%**	2	3	4	**47%**	2	3	4	**28%**	3	3	4	**50%**	2	3	4	**50%**
Requester View	Necessity	4	4	5	**80%**	4	4	5	**75%**	3	4	4	**51%**	3	4	5	**72%**	3	4	4	**72%**
	Accuracy	3	4	5	**70%**	3	4	5	**67%**	3	3	4	**47%**	3	4	5	**69%**	3	4	4	**69%**
	Costs	2	3	4	**40%**	3	3	4	**45%**	3	3	4	**44%**	3	3	4	**39%**	3	3	4	**39%**

Table 3. Survey Results. The cells include the quartiles (25%, median, 75%). N = 159 survey participants. 5-point Likert-scale from "strongly disagree" (1) to "strongly agree" (5). "Agree" is the percentage of people who answered 4 (agree) or strongly agree (5).

	Fairness	Offense	Benefit	Necessity	Accuracy	Costs
Qualification Test	**1**	**4**	**1**	**1**	**1**	4
Qualification Restriction	2	2	2	2	3	**1**
Majority Vote	5	1	5	5	5	**1**
Gold Standard	2	**4**	2	2	**1**	**1**
Validating Review	4	2	4	4	4	4
	1 = fair 5 = unfair	1 = offensive 5 = not offensive	1 = beneficial 5 = not beneficial	1 = required 5 = not required	1 = Higher quality 5 = Lower quality	1 = Lower cost 5 = Higher cost

Table 4. Ranking of Quality Assurance Mechanisms. For each category, the numbers indicate the relative position when comparing the five mechanisms. If non-parametric tests show no significant difference, two or more positions are grouped. Note: For "offense", less offensive is better.

The data shows that the survey participants are very well aware of the two QAMs that are openly visible on MTurk. QT (81%) and QR (86%) have been encountered "Sometimes," "Very Often," or "Always." The other three mechanisms that work in the background are less well known (MV 57%, VR 50%, GS 40%). *Negative experiences* with the mechanisms are common for all participants. About one third of the participants (QT 31%, QR 33%, MV 36%) remember having had problems with the mechanism at least sometimes. The results are lower for VR (19%) and GS (20%).

However, when looking at the combination of frequency and negative experience, the data shows that negative experiences are strongly associated with frequency of occurrence. In other words, many workers only realize that one of the background mechanisms is in use when they have a negative experience. If the survey participants are aware that a HIT on MTurk uses one of the QAMs, this has no influence for a majority of them. However, many workers are less likely to work on tasks using MV (40%) or VR (26%). Some workers (26%) also do not like to take a Qualification Test.

Again, these results are worse once the workers repeatedly encounter the mechanism or had negative experiences. If this is the case, they are significantly less inclined to work on future tasks given certain QAMs. These results suggest that poorly implemented QAMs can leave a long lasting negative impression.

Influence of demographics

After describing the overall results, we form groups of survey participants based on the demographic information and analyze the effect on the QAM ratings. Since many requesters are interested in finding the best trained workers and keeping them for their tasks (Ipeirotis, 2011), our goal is to understand the view of those workers that spend a lot of time on MTurk, have a high number of approved HITs, and a high HIT approval rate.

The results show that workers who spend a long time per week on MTurk (more than 20 hours, N=31) rank Majority Vote and Validating Review significantly less fair, less beneficial, and unnecessary. Thus, they are significantly less willing to work on tasks that use one of these two mechanisms. They also try to avoid gold standards. For workers with high quality results (self-reported HIT approval rate of 99% or higher; N = 67), the results are similar.

Overall, when the results are filtered according to experienced and non-experienced workers, the data shows that experienced workers realize that the "background" mechanisms have a negative impact in terms of the worker-view aspects. These experienced workers do not like to see their results compared to or evaluated by others.

The level of education had very little impact on the results. This is consistent with earlier work that suggests that the skills required for crowdsourcing are different from "formal education" and more specific capabilities like computer skills, language, or web navigation are required to perform tasks on MTurk (Schulze et al., 2011). The only interesting significant result is that workers with a Bachelor's degree or higher understand the necessity of the gold standard better.

DISCUSSION

Implications for theory

Our conceptual model contributes to theory in two ways. First, we have taken theory from related work in psychology, business, economics, affective science, etc. and applied it to the domain of online labor markets. The special characteristics of these labor markets is that the responsibility for matching tasks with adequate employees is mostly performed by the workers rather than Human Resources (Schulze, Krug, et al., 2012). Our work suggests that theory from traditional labor settings can be adapted to these new conditions and can contribute to further research in the worker view on these online labor markets. Second, we contribute to the area of statistical quality assurance in Human Computation. We show that the effect of the mechanism on the worker can be significant on the composition of the worker pool. This knowledge may be useful for the development of new and better QAMs, especially for approaches aiming at improving the initial data quality.

Implications for practice

Due to space limitations, only selected results could be mentioned in the previous chapter. Still, we have performed additional statistical analyses to test these implications. Based on the results of the study, we can formulate the following guidelines for platform operators and requesters:

Choice and Design of Mechanism. Approaches where the skills are tested before beginning the task (Qualification Test) or the worker population is restricted beforehand (Qualification Restriction) are generally accepted by the workers. Workers do not like Majority Vote which can be explained by the fact that the "correct" answer may be rejected if it disagrees with the "wrong" majority. We therefore strongly suggest that Majority Vote should not be the criterion to decide whether workers are paid or not. We would rather recommend that quality management and remuneration are viewed separately, i.e., that workers are still paid for all their work as long their rating is high enough (as opposed to punishing every single disagreement with the majority). Independently, an appropriate statistical approach can still be used to clean the results. If *Gold Standard* is used, it should be implemented carefully so that negative experiences (e.g., a correct answer is rejected) can be avoided for the experienced workers who sidestep tasks with this mechanism after negative experiences. *Validating Review* should ensure that the review process is thorough, i.e., that the chance of rejection of correct work is minimized.

Communication of Mechanism used. Requesters should clearly communicate the QAM used, the reason why it is used, and how the results will affect the workers in terms of payment and potential rejection. While this approach might discourage some workers from working on the task at all, it seems better than surprising honest workers with a feeling of being treated unfairly once a negative experience reveals the underlying mechanism. While we have not explicitly studied this potential effect, comments from the workers confirm that the exact design of a mechanism and the wording of explanations are important. This has been studied by Shaw et al. (2011) for a content analysis task.

Since our analysis was exploratory, all these findings should be tested again in a formal experiment. Depending on the task type, the complexity of a task, and the overall worker population, the implications for practice might be different.

CONCLUSION

Utilizing crowdsourcing marketplaces can be a convenient and cost-efficient way to recruit workers for human computation systems. Because of the open nature of most of these marketplaces, requesters need to utilize appropriate quality assurance mechanisms in order to validate the submitted results and to decide how much to pay to which worker. Previous research has mostly either focused on the statistical aspects of quality assurance, thus treating the worker as mere "noisy input," or dealt with alternative ways to motivate workers to submit results of higher quality. In contrast, we directly analyze the worker perception of well-known quality assurance mechanisms according to subjective and objective criteria. Based on theory from related areas, we develop a conceptual model and test it with a survey on Mechanical Turk. Our results show that the perception of the mechanisms is indeed different, especially with respect to the Majority Vote approach. Based on our results, we give the requesters on crowdsourcing markets the advice to integrate the worker view when selecting an appropriate mechanism.

As a first exploratory analysis of the research question, the results might not directly be applicable to any real world cases. Improving worker perception should only be one auxiliary means -besides other considerations like task type or task characteristics- when implementing an appropriate QAM. Related questions concerning the influence of QAMs on task performance are also not part of the analysis in this paper. In future work, the findings will therefore be tested in a controlled experiment that incorporates other metrics instead of just worker perception in order to analyze its importance more rigorously. Alternatively, the results can form the basis for a structural model that formally analyzes work intention based on the different aspects.

The presented study is not without limitations. The survey method used for the study might have led to a selection bias of the participants. It is unclear whether the population of the survey participants is equal to the population of the MTurk platform; and whether the quality assurance mechanisms used within the survey might have further biased the task selection. Since the MTurk platform is very open and empirical evidence shows that it attracts many dishonest workers and requesters, the results might not apply to other platforms with more control, i.e., external reliability might be limited; however, existing research indicates that rating scales are perceived differently by crowd workers (Riedl et al., 2013). To have a brief and concise model, we have limited it to three subjective and three objective aspects. We cannot be sure that these aspects are indeed the most important ones on MTurk. Future qualitative research, e.g., through informal interviews, could evaluate the robustness of the model. Finally, the research might benefit from being viewed from different theoretical lenses, e.g.m using principal-agent theory and related economic theories.

ACKNOWLEDGMENTS

The authors would like to thank Thomas Kutschker, Bachelor of Science (B.Sc.) in Business Informatics at the University of Mannheim. The survey data for this paper was collected as part of his Bachelor thesis. The authors would also like to thank the anonymous reviewers for the comprehensive feedback that greatly improved this paper and included various suggestions for future research.

REFERENCES

1. Aquino, K., Tripp, T. M., and Bies, R. J. 2006. "Getting even or moving on? Power, procedural justice, and types of offense as predictors of revenge, forgiveness, reconciliation, and avoidance in organizations," *Journal of Applied Psychology* (91:3), pp. 653–668.
2. Balkin, D. B., and Griffeth, R. W. 1993. "The determinants of employee benefits satisfaction," *Journal of Business and Psychology* (7:3), pp. 323–339.
3. Boone, H. N., Jr., and Boone, D. A. 2012. "Analyzing Likert Data," *The Journal of Extension (JOE)* (50:2).
4. Calhoun, C. 2004. "Subjectivity and emotion," In *Thinking about Feeling□: Contemporary Philosophers on Emotions*, Oxford University Press, pp. 107–121.
5. Chesbrough, H. 2003. *Open Innovation: The New Imperative for Creating and Profiting from Technology*, (1st ed), Mcgraw-Hill Professional.
6. Cornell, B., and Welch, I. 1996. "Culture, Information, and Screening Discrimination," *Journal of Political Economy* (104:3), pp. 542–71.

7. Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. "Are your participants gaming the system?," In *Proceedings of the 28th international conference on Human factors in computing systems (CHI'10)*, Atlanta, Georgia, USA, pp. 2399.
8. Gilliland, S. W. 1993. "The perceived fairness of selection systems: An organizational justice perspective," *The Academy of Management Review* (18:4), pp. 694–734.
9. Harris, M. M., Dworkin, J. B., and Park, J. 1990. "Preemployment screening procedures: How human resource managers perceive them," *Journal of Business and Psychology* (4:3), pp. 279–292.
10. Holmstrom, B., and Milgrom, P. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics & Organization* (7), pp. 24.
11. Howe, J. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, Crown Publishing Group.
12. Iles, P., Mabey, C., and Robertson, I. 1990. "HRM Practices and Employee Commitment: Possibilities, Pitfalls and Paradoxes," *British Journal of Management* (1:3), pp. 147–157.
13. Ipeirotis, P. 2011. "Does lack of reputation help the crowdsourcing industry? | A Computer Scientist in a Business School," http://www.behind-the-enemy-lines.com/2011/11/does-lack-of-reputation-help.html
14. Ipeirotis, P. G., Provost, F., and Wang, J. 2010. "Quality management on Amazon Mechanical Turk," In *Proceedings of the ACM SIGKDD Workshop on Human Computation,* Washington DC, ACM, pp. 64–67.
15. Jung, H. J., and Lease, M. 2011. "Improving Consensus Accuracy via Z-Score and Weighted Voting," In *The 3rd Human Computation Workshop (HCOMP 2011).*
16. Kaufmann, N., Schulze, T., and Veit, D. 2011. "More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk," In *AMCIS 2011 Proceedings*.
17. Kazai, G. 2010. "An Exploration of the Influence that Task Parameters have on the Performance of Crowds," Presented at the CrowdConf 2010, October 4, 2010, San Francisco, CA.
18. Kern, R., Bauer, C., Thies, H., and Satzger, G. 2010. "Validating results of human-based electronic services leveraging multiple reviewers," *AMCIS 2010 Proceedings*.
19. Kern, R., Thies, H., and Satzger, G. 2011. "Efficient Quality Management of Human-Based Electronic Services Leveraging Group Decision Making," *ECIS 2011 Proceedings*.
20. Kern, R., Zirpins, C., and Agarwal, S. 2009. "Managing Quality of Human-Based eServices," In *Service-Oriented Computing – ICSOC 2008 Workshops.* Lecture Notes in Computer Science, G. Feuerlicht and W. Lamersdorf (eds.), (Vol. 5472), Springer Berlin / Heidelberg, pp. 304–309.
21. Leimeister, J. M., Huber, M., Bretschneider, U., and Krcmar, H. 2009. "Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition," *Journal of Management Information Systems* (26:1), pp. 197–224.
22. Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. 2010. "Exploring iterative and parallel human computation processes," In *KDD-HCOMP'10,* Washington, DC, USA.
23. Lomas, P. 1994. *True & False Experience: The Human Element in Psychotherapy*, Transaction Publishers.
24. Macan, T. H., Avedon, M. J., Paese, M., and Smith, D. E. 1994. "The effects of applicants' reactions to cognitive ability tests and an assessment center," *Personnel Psychology* (47:4), pp. 715–738.
25. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. 2011. "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing," In *The 3rd Human Computation Workshop (HCOMP 2011).*
26. Quinn, A. J., and Bederson, B. B. 2011. "Human Computation: A Survey and Taxonomy of a Growing Field," In *CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.*
27. Riedl, C., Blohm, I., Leimeister, J. M., and Krcmar, H. 2013. "The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities," *International Journal of Electronic Commerce (IJEC)* (17:3), pp. 7–36.
28. Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. 2010. "Who are the crowdworkers?: shifting demographics in mechanical turk," In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, Atlanta, Georgia, USA*, ACM, pp. 2863–2872.

29. Ryan, A. M., and Ployhart, R. E. 2000. "Applicants' Perceptions of Selection Procedures and Decisions: A Critical Review and Agenda for the Future," *Journal of Management* (26:3), pp. 565–606.
30. Ryan, A. M., and Sackett, P. R. 1987. "Pre-employment honesty testing: Fakability, reactions of test takers, and company image," *Journal of Business and Psychology* (1:3), pp. 248–256.
31. Rzeszotarski, J. M., and Kittur, A. 2011. "Instrumenting the crowd: using implicit behavioral measures to predict task performance," In *Proceedings of the 24th annual ACM symposium on User interface software and technology (*UIST '11), New York, NY, USA: , ACM, pp. 13–22.
32. Sackett, P. R., and Wilk, S. L. 1994. "Within-group norming and other forms of score adjustment in preemployment testing," *American Psychologist* (49:11), pp. 929–954.
33. Schenk, E., and Guittard, C. 2011. "Towards a characterization of crowdsourcing practices," *Journal of Innovation Economics* (7:1), pp. 93–107.
34. Schulze, T., Indulska, M., Geiger, D., and Korthaus, A. 2012. "Idea assessment in open innovation: A state of practice," In: *20th European Conference on Information Systems (ECIS)*. Barcelona, Spain.
35. Schulze, T., Krug, S., and Schader, M. 2012. "Workers' Task Choice in Crowdsourcing and Human Computation Markets," In *International Conference on Information Systems (ICIS), Orlando, FL, USA*.
36. Schulze, T., Seedorf, S., Geiger, D., Kaufmann, N., and Schader, M. 2011. "Exploring Task Properties in Crowdsourcing - An Empirical Study on Mechanical Turk," In *19th European Conference on Information Systems (ECIS),* Helsinki, Finland.
37. Shaw, A. D., Horton, J. J., and Chen, D. L. 2011. "Designing incentives for inexpert human raters," In *Proceedings of the ACM 2011 conference on Computer supported cooperative work (*CSCW'11), New York, NY, USA, ACM, pp. 275–284.
38. Sorokin, A., and Forsyth, D. 2008. "Utility data annotation with Amazon Mechanical Turk," In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pp. 1 –8.
39. Sun, Y.-A., Roy, S., and Little, G. 2011. "Beyond Independent Agreement: A Tournament Selection Approach for Quality Assurance of Human Computation Tasks," In *The 3rd Human Computation Workshop (HCOMP 2011)*.
40. Tabibnia, G., Satpute, A. B., and Lieberman, M. D. 2008. "The Sunny Side of Fairness Preference for Fairness Activates Reward Circuitry (and Disregarding Unfairness Activates Self-Control Circuitry)," *Psychological Science* (19:4), pp. 339–347.
41. Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., Marin, D., and Simons, H. 2010. "Towards Building a High-Quality Workforce with Mechanical Turk," In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS), Whister.*
42. Weathington, B. L., and Tetrick, L. E. 2000. "Compensation or Right: An Analysis of Employee 'Fringe' Benefit Perception," *Employee Responsibilities and Rights Journal* (12:3), pp. 141–162.
43. Yampolskiy, R. V. 2012. "AI-Complete, AI-Hard, or AI-Easy: Classification of Problems in Artificial Intelligence," *Proceedings of the Twenty-third Midwest Artificial Intelligence and Cognitive Science Conference (MAICS-2012),* Cincinnati, Ohio, Omnipress, pp. 94–102.