Summer 10-6-2011

EXPLORING TASK PROPERTIES IN CROWDSOURCING – AN EMPIRICAL STUDY ON MECHANICAL TURK

Thimo Schulze

Stefan Seedorf

David Geiger

Nicolas Kaufmann

Martin Schader

Recommended Citation

Schulze, Thimo; Seedorf, Stefan; Geiger, David; Kaufmann, Nicolas; and Schader, Martin, "EXPLORING TASK PROPERTIES IN CROWDSOURCING – AN EMPIRICAL STUDY ON MECHANICAL TURK" (2011). *ECIS 2011 Proceedings*. 122.
http://aisel.aisnet.org/ecis2011/122

EXPLORING TASK PROPERTIES IN CROWDSOURCING – AN EMPIRICAL STUDY ON MECHANICAL TURK

Schulze, Thimo, University of Mannheim, Chair in Information Systems III, Schloss, 68131 Mannheim, Germany, schulze@wifo.uni-mannheim.de

Seedorf, Stefan, University of Mannheim, Chair in Information Systems III, Schloss, 68131 Mannheim, Germany, seedorf@wifo.uni-mannheim.de

Geiger, David, University of Mannheim, Chair in Information Systems III, Schloss, 68131 Mannheim, Germany, geiger@wifo.uni-mannheim.de

Kaufmann, Nicolas, mail@nicolas-kaufmann.de

Schader, Martin, University of Mannheim, Chair in Information Systems III, Schloss, 68131 Mannheim, Germany, martin.schader@uni-mannheim.de

Abstract

In the last years, crowdsourcing has emerged as a new approach for outsourcing work to a large number of human workers in the form of an open call. Amazon's Mechanical Turk (MTurk) enables requesters to efficiently distribute micro tasks to an unknown workforce which selects and processes them for small financial rewards. While worker behavior and demographics as well as task design and quality management have been studied in detail, more research is needed on the relationship between workers and task design. In this paper, we conduct a series of explorative studies on task properties on MTurk. First, we identify properties that may be relevant to workers' task selection through qualitative and quantitative preliminary studies. Second, we provide a quantitative survey with 345 participants. As a result, the task properties are ranked and set into relation with the workers' demographics and background. The analysis suggests that there is little influence of education level, age, and gender. Culture may influence the importance of bonuses, however. Based on the explorative data analysis, five hypotheses for future research are derived. This paper contributes to a better understanding of task choice and implies that other factors than demographics influence workers' task selection.

Keywords: Amazon Mechanical Turk, Cultural Differences, Survey, Crowdsourcing

1 Introduction

"Crowdsourcing," first mentioned by Howe (2006), can be defined as the act of taking a task once performed by the employees of a company and outsourcing it to a large, undefined group of people in an open call (Howe, 2008). The term has been used for a wide variety of phenomena and is related to areas like Open Innovation, Co-Creation, or User Generated Content. Recently, the area of "paid crowdsourcing" has gained a lot of momentum, with companies like CrowdFlower (www.crowdflower.com) and CloudCrowd (www.cloudcrowd.com) receiving big venture funding (Techcrunch.com, 2010a, 2010b). Frei (2009) defines paid crowdsourcing as using a technology intermediary for outsourcing paid work of all kinds to a large group of workers. Because of the dynamic scalability, paid crowdsourcing is often compared to cloud computing (Corney et al., 2009; Lenk et al., 2009).

Paid crowdsourcing on a large scale is enabled by platforms that allow requesters and workers to allocate resources. Amazon Mechanical Turk (www.mturk.com) is a market platform that gives organizations ("Requesters") the opportunity to get large amounts of work completed by a cost-effective, scalable, and potentially large number of disengaged workers ("Turkers"). Requesters break down jobs into micro tasks called HITs (Human Intelligence Tasks) which are selected and completed by human workers for a relatively small reward. Example tasks include image labeling, transcription, content categorization, and web research. However, this open nature of task allocation exposes the requester to serious problems regarding the quality of results. Some workers submit HITs by randomly selecting answers, submitting irrelevant text, etc., hoping to be paid for simply completing a task. Besides this inevitable "spam problem," reasons for bad results may include that workers did not understand the requested task or were simply not qualified to solve it.

Verifying the correctness of every submitted solution can often be as costly and time-consuming as performing the task itself (Ipeirotis et al., 2010). The prevalent solution to deal with these issues is the implementation of suitable quality management measures. A common approach is redundant assignment of tasks to multiple workers in combination with a subsequent comparison of the respective results. Another option is peer review where results from one worker are verified by others with a higher level of credibility (Kern et al., 2010). The resources invested into these measures can constitute a considerable overhead and diminish the efficiency of micro-task crowdsourcing.

Research has shown that the quality of task results can be substantially improved by choosing an adequate task design (Huang et al., 2010). Depending on the type and background of a task, its presentation may influence the result quality in two ways: First, a good and appropriate design facilitates the overall understanding of the job and therefore increases the chance of correct results (Khanna et al., 2010). Second, certain properties of the design may influence not only the overall attractiveness of the task but also the interest of specific groups of workers. Finding out what these properties are and how they correspond with demographics may enable organizations to target particularly qualified or motivated people, or, in other words, attract the right crowd for a given job. The objective of this contribution is to identify some of these potentially relevant properties by conducting studies with the workers themselves. In particular, we try to answer the following research questions:

- Which task properties are important for crowdsourcing workers and influence task selection?
- Does the personal situation or demographic background of workers impact the task properties they perceive as being more important?

Significant work on demographics and the worker perspective in crowdsourcing as well as task design and quality management has been done, but the relation of both areas has not been studied in detail. We therefore follow an explorative research approach using qualitative and quantitative methods. As a prerequisite for future research directions, the goal of this paper is to get more comprehensive insights into paid crowdsourcing.

The paper is organized into six sections. After introducing the related work in the next part, section three explains the explorative bottom-up approach of the paper and how the original idea was refined using a series of preliminary studies. Section four describes the design and analysis of the main survey. The corresponding results are evaluated and discussed in section five. We conclude the paper with a summary of our findings and an outlook on future work.

2 Related Work

Related work for our paper comes from two areas of crowdsourcing research. First, our paper builds on the behavior and demographics of the workers on crowdsourcing platforms and the way they use the platforms. Second, good task design and quality management mechanisms which are important for sound use of crowdsourcing platforms are presented.

2.1 Crowdsourcing Demographics and the Worker Perspective

Ross et al. (2010) and Ipeirotis (2010) give a comprehensive overview of the demographics of workers on MTurk. We will discuss their results in more detail in section 4.3. Silberman et al. (2010) explain challenges that workers face on MTurk including fraudulent requesters, privacy problems, and tasks with unclear instructions. They list tools that help workers to avoid these drawbacks. However, Horton (2010) finds that requesters on MTurk do treat workers as honestly and fairly as other employers from their home country. Khanna et al. (2010) focus specifically on low-income workers from India and conclude that these workers require simplified user interfaces and descriptions to participate on platforms like MTurk.

Downs et al. (2010) perform a large scale study on MTurk and find that some respondents participate for quick cash and do not respond conscientiously. They also identified demographic differences, e.g., that young men seem to be more likely trying to game the system while professionals, students, and non-workers take tests more seriously. Chandler and Kapelner (2010) conduct a natural field experiment to explore the relationship between the meaningfulness of tasks and the willingness of workers to work on the tasks. They see no increase in quality for meaningful tasks, as opposed to quantity. Chilton et al. (2010) use a multi-method approach to research how workers search for tasks on MTurk. Based on a survey and a high frequency data scrape, they conclude that workers mostly select tasks that were most recently posted or with most available instances left.

2.2 Task Design and Quality Management

Huang et al. (2010) aim towards finding optimal task parameters to maximize the number of quality image labels given time and budget constraints. Ipeirotis et al. (2010) propose a new estimation approach for blocking low performance workers and spammers from performing further work while separating true errors from biased answers. Kern et al. (2010) introduce a majority review approach where workers review and validate the results submitted by others. A similar approach is also used by CloudCrowd (www.cloudcrowd.com). Raykar et al. (2010) propose a probabilistic approach for dealing with noisy labels that shows better results than the simple majority voting baseline.

Recently, the economics of crowdsourcing have been studied as well. Horton and Chilton (2010) introduce a method for calculating the reservation wage of workers. Mason and Watts (2009) discover that increased payments increase only the quantity of the work but not the quality. Further, they found out that the compensation scheme influences the output considerably.

2.3 Surveys on Amazon Mechanical Turk

Mechanical Turk has been used for experiments and surveys in various domains. Kittur et al. (2008) conclude that MTurk is applicable for a variety of user study tasks because hundreds of participants can be recruited in a short timeframe for marginal costs. Paolacci et al. (2010) compare the results of experiments conducted on MTurk to current methods of recruiting subjects at universities or Internet boards. They conclude that MTurk workers pay attention to directives at least as much as subjects from traditional sources. Heer and Bostock (2010) also conclude that the MTurk contributions are viable and a high level of validity can be expected. For the design of our survey, we used the guidelines and best practices given by these authors and additional sources.

3 Preliminary Explorative Studies

Since we could not identify any suitable theories that explore the different aspects of task properties or link worker demographics to these properties, we decided to analyze the research questions by means of an explorative approach. In order to obtain a comprehensive understanding of the relevant dimensions, we chose an incremental bottom-up approach starting from the workers' perceptions. The preliminary studies included various qualitative and quantitative studies.

3.1 Approach for Preliminary Studies

(1) We started to explore the research questions by asking the workers on Mechanical Turk an open question about their five most important properties in task descriptions. In about 24 hours, we collected 50 answers for 0.05 USD each. We then classified these properties into 15 different general properties.

(2) Next, we provided the list of these 15 aggregated properties to 100 workers and asked them to select the five most important properties. We also gave them the option to list additional important properties in the comments.

The selected properties from these two studies showed some noticeable differences. For example, Payment/Reward was only mentioned in 33% of the cases when asked openly, but was selected in 60% of the cases when the workers were asked to select out of predefined options. This is consistent with studies of Shaw (2010) which suggest that workers on Mechanical Turk give mixed results about the importance of money depending on whether they are asked directly or indirectly. While we first wanted to focus on the descriptions of the tasks, workers told us in the comments that general properties like title, requester, number of HITs available, and payment are more important to them. Therefore, the property list was extended and modified to include a total of 21 properties.

(3) Subsequently, we asked the workers to rate the importance of each of these 21 properties on a 5-point Likert scale. (4) In addition, we formulated statements for both extremes for every property (e.g., "I only work on HITs that sound interesting, enjoyable, or fun." and "I do not care if a HIT sounds interesting, enjoyable or fun."). We then asked the workers to choose the statement they would most agree with. Again, the results were considerably different for these two ways of asking the questions. The results suggest that there might be differences between properties that are really important for the workers to start working on a task at all, and other properties that are valued as nice to have, but do not generally influence the workers' selection decision that much.

(5) As a final qualitative survey, we provided the complete property list to experienced MTurk workers (more than 10,000 completed HITs) and offered a high bonus for new and important properties that we might have missed. We received ten answers within one day that mainly extended existing properties.

The first three preliminary studies were posted on MTurk via Crowdflower (www.crowdflower.com) for reasons of interface usability. The remaining two studies were posted directly on MTurk. Altogether, responses from 410 participants were collected in the five preliminary studies for a total payment of US$38.30 including fees.

3.2 Implications from Preliminary Studies

Since we used a bottom-up approach, the preliminary analyses were conducted in a broad manner, including even aspects that should be self-evident. While our explorative approach does not explicitly test hypotheses, some results seem to remain consistent, independently of the different ways the workers were questioned.

Clear Instructions. Workers want the instructions to be clear and complete. Especially for short and simple tasks, it is extremely important that they know what to do and that the description does not contain ambiguities about what output is expected from them.

Genuine / No scam. MTurk is an open marketplace where everyone can post HITs. While terms clearly state that HITs involving activities like spamming, direct marketing, or fraud are not allowed, some of these illegitimate HITs are still posted on the site, occasionally. Our research shows that a vast majority of workers only wants legitimate work and not work on HITs that they refer to as "scam."

No need to leave MTurk Platform. Many workers do not want to leave the site to work on tasks since they had bad experiences with malicious webpages. Therefore, tasks should be designed to be solvable within the crowdsourcing platform whenever possible.

4 Task Properties – Main Survey

With these five preliminary studies, we were able to compose a comprehensive selection of task properties that might influence the task selection of different workers. To analyze the importance of these properties and the impact of demographics and personal background on them, we designed a large-scale quantitative survey.

4.1 Excluding Properties Identified in Preliminary Studies

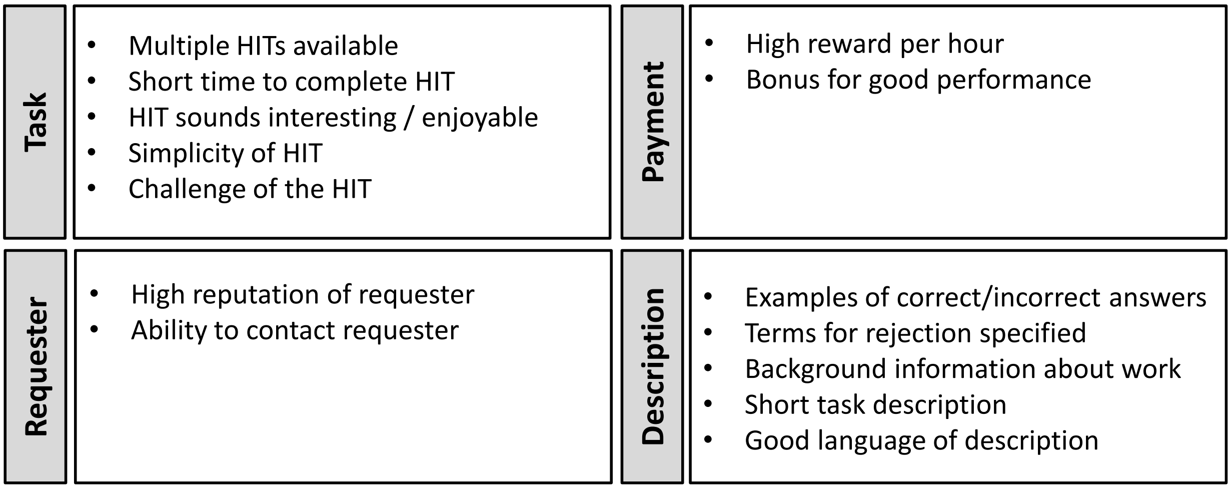
To keep the survey short, it was necessary to restrict the number of included task properties. In the preliminary studies we used rather vague wording whether workers wanted the reward per single HIT to be high or low. Naturally, workers preferred high rewards. However, we also observed a correlation between *High reward per HIT* and *Short time to complete a HIT*. On the one hand, workers commented that small tasks paying "just a penny" are often associated with low hourly wages. On the other hand, they are reluctant to work on a single task for a long time if the complete effort might be rejected because of small mistakes.

In a news article, the crowdsourcing researcher Luis von Ahn suggests that the optimal size of individual tasks might be different depending on the task type. He suggests that researchers need to develop a better theoretical understanding of the relationship between task sizes, cost, and time (Hardesty, 2010). We also suggest that questions regarding task size and reward per task can better be analyzed by other research methods (e.g., natural experiments); therefore we excluded these questions from the survey. Huang et al. (2010) and Kazai (2010) perform first experiments in that direction by varying task and reward size for specific tasks.

We also exclude two properties that were identified in the preliminary studies from further research because they are very specific to the Mechanical Turk platform: *Time allotted*, which is defined as the time the workers have to complete a HIT starting from the moment they accept it, is not important to most workers as long as the task can be comfortably completed in that time. Also, the *requirement to take a qualification test* is rather a quality management mechanism or entry barrier that goes beyond task properties.

4.2 Survey Design

After the analysis, our final survey included 14 properties of tasks on paid crowdsourcing platforms. All properties are not directly dependent on a specific task type or task size. The properties can be structured into the four categories task, payment, description, and requester (as depicted in Figure 1).



Task

- Multiple HITs available
- Short time to complete HIT
- HIT sounds interesting / enjoyable
- Simplicity of HIT
- Challenge of the HIT

Payment

- High reward per hour
- Bonus for good performance

Requester

- High reputation of requester
- Ability to contact requester

Description

- Examples of correct/incorrect answers
- Terms for rejection specified
- Background information about work
- Short task description
- Good language of description

Figure 1: Categories of task properties

Survey Design. The survey consists of three parts. (1) For the first part, the properties were transformed into statements, all starting with “I only work on HITs …,” e.g., “I only work on HITs that are challenging.” or “I only work on HITs that sound interesting or enjoyable.” We asked the workers to state how much they agree with these statements using a finer grained 7-point Likert scale (Strongly disagree, Moderately disagree, Slightly disagree, Neutral, Slightly agree, Moderately agree, Strongly agree). The order of the 14 questions was randomized. (2) To reduce the risk of cross-cultural bias the rating was complemented by a ranking of alternatives (Harzing et al., 2009). We presented the workers with a randomized list of the 14 properties. They were then asked to select and rank the five most important properties. For usability reasons, we used the drag and drop interface of SurveyGizmo for this task. (3) Finally, we collected demographic information from the workers (gender, age, country, level of education, the importance of MTurk money for the worker, the time they have been on MTurk, and the time they work on MTurk per week). We used elements similar to the related surveys of Ross et al. (2010) and Ipeirotis (2010).

Quality Management Measurements. As MTurk research shows issues with random answers on surveys, measures are needed to ensure quality (Kapelner and Chandler, 2010; Downs et al., 2010). Due to the exploratory character of the survey and the goal to keep it short and concise, we decided not to ask multiple questions for each construct. We verified quality by inserting two “test questions” in the survey, instead. To avoid offending honest workers, the reason for these test questions was explained in detail within the instructions. The questions were worded like “I only work on HITs. Attention. This is a test question. Please select "Moderately disagree" here.”

Distribution and Pricing. A draft of the survey was analyzed by five experts. Due to their feedback, the wording of some questions was changed to avoid any ambiguities. Since the usual payment on MTurk is about US$1-2 per hour (Horton and Chilton, 2010), we decided to pay the workers US$0.15 to complete the survey, with an estimated completion time of 6 min, resulting in a pay-out of US$1.5 per hour. The survey itself was designed with SurveyGizmo (www.surveygizmo.com). The task on MTurk contained a link to the survey and a text box to paste the completion code that we provided at the end of the survey.

4.3 Survey Data

We launched the survey on a Thursday morning in November 2010. After five days, we had collected 447 responses. MTurk calculated an average response time of 5 minutes and an effective hourly rate of $1.65. Unfortunately, 102 workers answered at least one of the two "test questions" incorrectly and were excluded from the results. 79 of the invalid answers could be geo-located to respondents from India. Our score of 77% of valid responses is consistent with results by (Downs et al., 2010) that see a correct response rate between 61% and 88%, depending on the difficulty of the spam detecting measure. Therefore, we paid a total of US$52.05 to the 345 workers who provided valid data.

The distribution of the participants' demographics is similar to other studies on MTurk. The 345 workers who provided valid data come from 31 countries. 34.8% are based in the USA, 50.4% are from India and 14.8% from other countries. This is a 14% higher share of Indians compared to Ross et al. (2010) and Ipeirotis (2010). However, this divergence is clearly explained by Ross et al. (2010) who find out that the worker population has been shifting towards a higher Indian participation in 2009. 49.3% of the workers are male and 50.7% female. 71% of the respondents are younger than 35 years of age, 28.4% are between 35 and 65 of age and only two are older than 65 years. Indian participants tend to be male (n=110) or young (age<35; n=138), whereas the majority of US participants are female (n=81).

Furthermore, various education levels are represented among the participants. 50.1% hold an Associate's or Bachelor's degree, 18.6% hold a Master's degree and 3.8% have a PhD or Professional degree. For 27.5% of the workers the earned money on MTurk is irrelevant or does not materially change their circumstances. For 39.4% of the participants it can be a nice extra and 33% responded that it is sometimes or always necessary. Ross et al. (2010) notice that self-selection can be a problem with surveys on MTurk. The results may be biased towards workers who enjoy doing surveys and therefore generally prefer more enjoyable tasks. But since the analysis of the demographics is in line with Ross et al. (2010) and Ipeirotis (2010), it can be assumed that the respondents represent a cross-section of the MTurk population.

4.4 Data Analysis Methods

The data analysis consisted of several stages and was facilitated by PAWS (SPSS) Statistics 18 and R 2.12.2. (1) First, we used the Likert ratings to calculate the averages of the task properties. (2) The results from the ranking section of the survey were determined by calculating the averages of their relative importance by assigning between 5 points (1st rank) and 1 point (5th rank) to the respective property. Both the ratings and the ranking were also determined separately for the USA, India, and other countries.

(3) Descriptive statistics and (4) a correlation analysis using Kendall's Tau for ordinal scales were conducted and analyzed in order to identify candidates for further statistical analysis. (5) Since correlation is an appropriate measure for some variables only, we grouped variables based on our explorative analysis (for example, the group of MTurkers who spend more than 8 hours a week on the platform) and applied non-parametric tests[1].

(6) To find patterns in the data, we performed a principal components analysis on the 14 properties[2] as well as (7) a cluster analysis. As the cluster analysis only provided clusters of low quality, we focused

[1] Since standard variance analyses like the ANOVA require normal distributions which could not be affirmed for all properties, we decided to mainly rely on nonparametric tests like the Mann-Whitney U or the Kruskal-Wallis test (H test). These are also reliable in the case of non-normality.

[2] We invoked R's function `principal` to calculate rotated principal components (varimax was used for rotation). The three component solution was chosen because a plot of the eigenvalues of the data matrix shows a clear "elbow" for three components.

on the principal components. (8) Based on the results, we later standardized the raw data and repeated some of the tests on the transformed data.

5 Results

5.1 General Results

The average ratings of the 14 task properties are shown in Table 1. The respondents agreed most with HITs that sound interesting/enjoyable, have good language of description, and return a high reward per hour. Further, examples of correct and incorrect answers, short description, and simplicity of the task were perceived desirable. Different results were observed for the importance ranking. Interesting HITs and high reward per hour are ranked highest as well, whereas good language was perceived less important (7th) compared to the simplicity of the HIT which is ranked 3rd by the workers, overall.

The challenge of the HIT, the ability to contact the requester and a bonus for good performance receive less than 3.5 on the Likert scale. However, a bonus for good performance is considered more important compared to other properties (6th rank). Ability to contact requester, bonus for good performance, and challenge of the HIT were ranked lowest.

Task Property	**Av. Rating**	**Av. Rating by Country**			**Rank**	**Rank by Country**		
		USA	**India**	**Other**		USA	**India**	**Other**
HIT sounds interesting / enjoyable	**5.21**	**4.93**	**5.33**	**5.45**	**2**	**1**	**3**	**2**
Good language of description	**4.94**	**4.82**	**5.18**	4.39	7	6	8	11
High reward per hour	**4.70**	**4.18**	**5.16**	**4.39**	**1**	**2**	**1**	**1**
Examples of correct/incorrect answers	**4.62**	4.04	**5.03**	**4.59**	8	10	7	10
Short task description	**4.59**	**4.17**	4.84	**4.76**	11	8	10	9
Simplicity of HIT	4.54	3.84	**4.96**	**4.73**	**3**	**3**	**2**	**3**
Terms for rejection specified	4.28	4.03	4.32	4.69	10	9	9	6
Short time to complete HIT	4.08	3.51	4.52	3.94	**4**	**5**	6	**4**
High reputation of requester	3.99	**4.08**	3.99	3.75	9	**4**	12	8
Background information about work	3.97	3.44	4.43	3.63	13	13	13	13
Multiple HITs available	3.80	2.88	4.59	3.25	**5**	7	**5**	7
Challenge of the HIT	3.47	2.83	4.02	3.14	12	12	11	12
Ability to contact requester	3.43	3.10	3.73	3.18	14	14	14	14
Bonus for good performance	3.33	2.40	4.07	3.00	6	11	**4**	**5**

*Five highest values have been highlighted in every column

Table 1: Rating and ranking of task properties

Cultural preferences. Since Indians seem to agree more with most properties, standardization was performed later in section 5.3. Therefore, only differences in the ranking are discussed here. Americans and Indians currently represent the two largest groups on MTurk and also in our survey. Other nationalities were grouped together due to small sample sizes. The ranking of task properties for both groups showed that the three most important properties are the same although they appear with varying priorities. Results indicate that the reputation of the requester is perceived important by the American respondents (4th rank), while it is not important for the Indian respondents (11th rank). On the other hand, Indian workers ranked bonus for good performance as the 4th most important task property. The American workers both ranked and rated this property much lower (11th rank). The question whether the cultural background is the influencing factor in these cases cannot be conclusively decided from the available data.

5.2 Principal Components Analysis

The three components that were identified in the principal components analysis are shown in Table 2. Based on the component loadings, we grouped the properties into categories and assigned descriptive titles to them.

Task Property	**Rotated Component**		
	1	**2**	**3**
Short task description	0.70	0.16	
Short time to complete HIT	0.60		0.32
Simplicity of HIT	0.78		
High reward per hour	0.50	0.16	0.30
Examples of correct/incorrect answers	0.42	0.59	0.24
Terms for rejection specified	0.21	0.60	
Good language of description	0.21	0.69	
High reputation of requester		0.69	0.22
Ability to contact requester		0.63	0.41
Multiple HITs available	0.48	0.15	0.54
HIT sounds interesting / enjoyable	0.23	0.14	0.52
Challenge of the HIT		0.10	0.83
Bonus for good performance	0.46	0.28	0.50
Background information about work		0.45	0.48

Table 2: Principal components analysis of the task properties

Quick Profit Jobbers. The properties with the highest loadings on *component 1* are the simplicity of the task, a short time needed for task completion, a short task description, and a comparatively high reward per hour. Workers who agree with this component can therefore be seen as the *Quick Profit Jobbers*.

Informed Workers. The highest loadings on *component 2* are the ability to get in personal contact with the requester, a good reputation of the requester, given examples for answers, the use of good language in the task description, and clearly specified terms for rejection of work. As this component seems to cover tasks that allow the worker to make a good personal impression on the requester, workers rating this component highly are *Informed Workers*.

Challenge Seekers. In c*omponent 3*, the challenge of the task has the most noticeable impact. The component also has high loadings for multiple tasks that are interesting, give background information and offer a bonus for good performance. We therefore classify workers valuing this component high as *Challenge Seekers*.

Agreement to the three components is not mutually exclusive, thus, there are some workers that agree with more than one component or none of them.

5.3 Standardization of Data

Motivation for Standardization. The results from Table 1 suggest that there is a significant difference of answering patterns between the cultural groups. A Kruskal-Wallis test[3] highlights that participants from India show a significantly higher voting attitude for 12 of the 14 properties, with no significant differences for the other two. Overall, the average Likert ratings among Indians (average of 4.58) are significantly higher compared to Americans (3.73) and other countries (4.06). Since participants from India are well-represented in the survey (174 of 345 participants), the influence of their over- or under-representation in certain groups of demographics could have a major influence on the results.

The higher ratings might be due to cultural differences in the response style, which is a common difficulty in cross-cultural research and is usually differentiated into two kinds of bias (Harzing et al., 2009): Acquiescent response style (ARS) refers to a general tendency to agree with a question and, thus, results in a bias towards the positive end of the rating scale. Extreme response style (ERS), respectively, refers to the tendency of preferring extreme responses on rating scales. The academic literature contains plenty of evidence for such response bias across a variety of countries including India (Johnson et al., 2005). Despite employing a 7-point Likert scale, as recommended by Harzing et al. (2009), the raw data still showed a significant ARS bias.

Standardization Method. Therefore, we decided to apply a widely used standardization method. According to Fisher and Milfont (2010), within-cultural standardization via Z-transformation is the best solution to overcome ARS. The resulting standardized values have a mean of 0 and a standard deviation of 1 over each cultural group. The value shows the position of a specific participant relative to all other participants in the same group. Following the transformation, a Kruskal-Wallis test on the standardized Z-values reveals that the origin of a participant no longer statistically influences the item rankings. Thus, cross-cultural differences are removed from the standardized data.

Results from standardized data. The standardized data set shows some interesting patterns. For example, workers spending more than 8h per week on Mechanical Turk show significantly higher values for items that have to do with “cost-efficient” working on tasks (High reward (p=0.011), bonus (p=0.012) and the possibility to work on similar tasks in the same task group (p=0.000)) as well as for the two “requester”-related items reputation (p=0.002) and contact possibility (p=0.001). Our analysis of the standardized data was the basis for the formulation of the hypotheses described in the next section. Due to the variety of explorative analysis methods used, not all observations that lead to these hypotheses can be described in this section because of space limitations.

6 Discussion of Results

Based on the results from the explorative data analysis, we formulate hypotheses that can be interesting directions for further research.

***H1**: The level of education does not have an effect on task property preferences.*

In every analysis method, we could not observe a connection between the highest level of education and any of the properties (except challenge of the HIT which is rated higher for workers with a Bachelor’s degree). A reason could be that skills learned from formal education are not required on MTurk where the required skill level is low overall and other capabilities like computer skills, language, or web navigation are required. A qualitative study asking open questions to MTurk workers

[3] The Kruskal-Wallis test – also known as H-test – is a nonparametric test that tests the hypothesis, that there is no difference between the means of two independent samples. Because it is based on ranks, the assumption of normality is not necessary (Kruskal and Wallis, 1952). This test was chosen because normal distribution cannot be guaranteed at this point of data analysis.

or an experiment could be used to study this aspect further. A related aspect is to determine whether the quality of results is higher for highly educated workers.

***H2**: Full time MTurk workers appreciate reputable requesters that communicate with workers.*

The data shows that workers who spend a long time per week (more than 8 hours, and especially more than 20 hours) emphasize the need for trustworthy requesters with a high reputation. Because of the consequences involved with being rejected or blocked wrongfully, they try to establish long time work relationships with a few honest requesters. Further research is needed to analyze whether good communication with workers can improve results and reduce cost and also to investigate what motivates workers to perform quality work over a long time.

***H3:** Indian workers are more likely to select a task on MTurk if they can achieve a bonus for good performance.*

Most tasks on MTurk currently only offer a fixed payment for every submission that is not rejected. Bonus payments for correct or above average results are not common. However, our results suggest that Indian workers on our sample prefer receiving a bonus as an affirmation of good work. Further research is needed to analyze whether bonus payments can improve result quality, enhance long term motivation of workers; and whether the effect varies for different cultures or task types. Bonuses could potentially also reduce the prevalent spam problem since submitting results of low quality would no longer be profitable.

***H4:** Aspects like behavior or motivational background play a more important role for preferences of crowdsourcing workers than demographics.*

In our data analysis, little effects could be observed for the demographic values age, gender, education, and period on MTurk. Nevertheless, the principal components analysis resulted in three distinct groups. This could suggest that other aspects like behavior or motivation play a more important role. The properties that are highly ranked correspond quite well with the job characteristic model by Hackman and Oldham (1980). Further research is needed to examine whether the motivation of Mturk workers corresponds with that of industry workers or not.

***H5**: Task Descriptions have to be designed differently for varying task types and worker groups.*

The component groups and the correlation analysis suggest that extent and complexity need to be adapted depending on task type and worker group. For example, short tasks with multiple HITs seem to require short and concise descriptions while comprehensive tasks like surveys can profit from extensive personal background information. These influences have not been studied in detail yet.

7 Conclusion and Future Work

Crowdsourcing platforms are an emerging way to outsource large amounts of labor to a distributed workforce. Companies can utilize platforms like Amazon Mechanical Turk to quickly solve problems that cannot adequately be solved in-house. While many studies have been conducted on various aspects of worker demographics and behavior as well as task design and quality management, the interrelation of those two areas has not been thoroughly studied yet. We approach this research question by analyzing which task properties are important for crowdsourcing workers and how this is influenced by demographic backgrounds. Our study combines qualitative and quantitative explorative research methods.

Our results contribute to a better understanding of the relevant task properties. We conclude that many workers like multiple, short, and simple tasks while others prefer comprehensive task descriptions. Americans seem to prefer reputable requesters while Indians might want to be rewarded by bonus payments. These findings on the workers' view could help requesters to design tasks better in order to attracting the right crowd for a given job and generating higher quality results. Based on our analysis, we identified interesting research directions for future work and formulated five hypotheses about task design, motivation, and cultural background. In a next step, these hypotheses have to be tested using other research methods like qualitative studies or natural experiments.

References

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010). Task search in a human computation market. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 1-9), Washington DC.

Corney, J. R., Torres-Sanchez, C., Jagadeesan, A. P., & Regli, W. C. (2009). Outsourcing labour to the cloud. International Journal of Innovation and Sustainable Development, Vol. 4 (No. 4), pp. 294-313.

CRAN. The Comprehensive R Archive Network. http://cran.r-project.org/

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Proceedings of the 28th international conference on Human factors in computing systems – CHI'10, Atlanta, Georgia, USA.

Fisher, R., & Milfont, T. L. (2010). Standardization in psychological research. International Journal of Psychological Research, 3(2), pp. 88-96.

Frei, B. (2009, September 15). Paid Crowdsourcing - Current State & Progress toward Mainstream Business Use. Retrieved from http://www.smartsheet.com/files/haymaker/Paid%20Crowdsourcing%20Sept%202009%20-%20Release%20Version%20-%20Smartsheet.pdf

Hackman, J., & Oldham, G. R. (1980). Work redesign. Addison-Wesley, Reading Mass.

Hardesty, L. (2010). Programming crowds. Retrieved from http://web.mit.edu/newsoffice/2010/programming-crowds-1027.html

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., Espejo, A., et al. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? International Business Review, 18(4), pp. 417-432.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception. Proceedings of the 28th international conference on Human factors in computing systems – CHI'10, Atlanta, Georgia, USA.

Horton, J. (2010). The Condition of the Turking Class: Are Online Employers Fair and Honest? Retrieved from http://arxiv.org/abs/1001.1172

Horton, J., & Chilton, L. B. (2010). The labor economics of paid crowdsourcing. Proceedings of the 11th ACM conference on Electronic commerce (pp. 209-218). Cambridge, Massachusetts, USA

Howe, J. (2006). The Rise of Crowdsourcing. Wired, 14(6). Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

Howe, J. (2008). Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Publishing Group.

Huang, E., Zhang, H., Parkes, D. C., Gajos, K. Z., & Chen, Y. (2010). Toward automatic task design: a progress report. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 77-85). Washington DC.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. Working Paper. Retrieved from http://archive.nyu.edu/handle/2451/29585

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on Amazon Mechanical Turk. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 64-67). Washington DC.

Johnson, T., Kulesa, P., Llc, I., Cho, Y. I., & Shavitt, S. (2005). The Relation Between Culture and Response Styles. Journal of Cross-Cultural Psychology, 36(2), pp. 264 -277.

Kapelner, A., & Chandler, D. (2010). Preventing Satisficing in Online Surveys: A "Kapcha" to Ensure Higher Quality Data. CrowdConf 2010, October 4, 2010, San Francisco, CA.

Kazai, G. (2010). An Exploration of the Influence that Task Parameters have on the Performance of Crowds. CrowdConf 2010, October 4, 2010, San Francisco, CA.

Kern, R., Bauer, C., Thies, H., & Satzger, G. (2010). Validating results of human-based electronic services leveraging multiple reviewers. AMCIS 2010 Proceedings. Paper 525. http://aisel.aisnet.org/amcis2010/525

Khanna, S., Ratan, A., Davis, J., & Thies, W. (2010). Evaluating and Improving the Usability of Mechanical Turk for Low-Income Workers in India. ACM DEV'10, December 17–18, 2010, London, United Kingdom.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI'08, Florence, Italy.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association, 47 (260), pp. 583-621.

Lenk, A., Klems, M., Nimis, J., Tai, S., & Sandholm, T. (2009). What's inside the Cloud? An architectural map of the Cloud landscape. Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing (pp. 23-31). IEEE Computer Society.

Mason, W., & Watts, D. J. (2009). Financial incentives and the "performance of crowds." Proceedings of the ACM SIGKDD Workshop on Human Computation – HCOMP'09. Paris, France.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. Judgment and Decision Making, Vol. 5 (No. 5), pp. 411-419.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., Moy, L., et al. (2010). Learning From Crowds. Journal of Machine Learning Research 11 (2010), pp. 1297-1322.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the Crowdworkers? Shifting Demographics in Mechanical Turk. Proceedings of the 28th international conference extended abstracts on Human factors in computing systems – CHI EA '10 (pp. 2863-2872). Atlanta, Georgia, USA.

Shaw, A. (2010). The CrowdFlower Blog - For love or for money? A list experiment on the motivations behind crowdsourcing work. Retrieved from http://blog.crowdflower.com/2010/08/for-love-or-for-money-a-list-experiment-on-the-motivations-behind-crowdsourcing-work/#more-931

Silberman, M. S., Ross, J., Irani, L., & Tomlinson, B. (2010). Sellers' problems in human computation markets. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 18-21). Washington DC.

Techcrunch.com. (2010a). CrowdFlower Raises $5 Million For Cloud Sourced Labor. Retrieved from http://techcrunch.com/2010/01/20/crowdflower-raises-5-million-for-cloud-sourced-labor/

Techcrunch.com. (2010b). CloudCrowd Raises $5.1 Million To Outsource Labor To The Cloud. Retrieved from http://techcrunch.com/2010/08/13/cloudcrowd-raises-5-1-million-to-outsource-labor-to-the-cloud/