Crowd IQ: Measuring the Intelligence of Crowdsourcing Platforms

Michal Kosinski
The Psychometrics Centre,
University of Cambridge, UK
mk583@cam.ac.uk

Yoram Bachrach
Microsoft Research,
Cambridge, UK
yobach@microsoft.com

Gjergji Kasneci
Microsoft Research,
Cambridge, UK
gjergjik@microsoft.com

Jurgen Van-Gael
Microsoft Research,
Cambridge, UK
jurgen@microsoft.com

Thore Graepel
Microsoft Research,
Cambridge, UK
thoreg@microsoft.com

ABSTRACT

We measure crowdsourcing performance based on a standard IQ questionnaire, and examine Amazon's Mechanical Turk (AMT) performance under different conditions. These include variations of the payment amount offered, the way incorrect responses affect workers' reputations, threshold reputation scores of participating AMT workers, and the number of workers per task. We show that crowds composed of workers of high reputation achieve higher performance than low reputation crowds, and the effect of the amount of payment is non-monotone—both paying too much and too little affects performance. Furthermore, higher performance is achieved when the task is designed such that incorrect responses can decrease workers' reputation scores. Using majority vote to aggregate multiple responses to the same task can significantly improve performance, which can be further boosted by dynamically allocating workers to tasks in order to break ties.

ACM Classification Keywords

H.4 Information Systems Applications: Miscellaneous

General Terms

Algorithms, Economics

Author Keywords

Crowdsourcing, Psychometrics, Incentive Schemes

INTRODUCTION

Consider a task relying heavily on mental abilities, such as solving an IQ test. Who would you expect to perform better: an average job applicant, or a small crowd composed of anonymous people paid a few cents for their work? Many would think that the single well-motivated individual should obtain a better or at least a comparable score. We show that even a small crowd can do better on an IQ test than 99% of the general population, and can also perform the task much faster.

The collective intelligence of crowds can be used to solve a wide range of tasks. Well known examples of platforms using the crowds' labour to solve complex tasks include Wikipedia, Yahoo! Answers, and various prediction markets [1, 10, 43]. Similarly, rather than relying exclusively on the labour of their own employees, institutions are using *crowdsourcing* to carry out business tasks and obtain information using one of the *crowdsourcing marketplaces*, such as Amazon Mechanical Turk[1] (AMT), Taskcn[2] and Crowd-Flower[3]. These marketplaces connect *workers*, interested in selling their labour, with *requesters* seeking crowds to solve their tasks.

Requesters split their problems into single tasks, so-called Human Intelligence Tasks[4] (HITs), and offer rewards to workers for solving them. Crowdsourcing markets offer great opportunities for both the requesters and the workers. They allow workers to easily access a large pool of jobs globally, and work from the comfort of their own homes. On the other hand, requesters gain instant access to very competitively priced labour, which can be quickly obtained even for a time-critical task.

Typical crowdsourced tasks include filling surveys, labelling items (such as images or descriptions) or populating databases. More sophisticated HITs may involve developing product descriptions, analysing data, translating short passages of text, or even writing press articles on a given subject. In our study, a worker's task was to answer a question from an IQ test.

Current implementations of crowdsourcing suffer from certain limitations and disadvantages and to use them effectively, one must devise appropriate designs for the tasks at hand. Different tasks may require solutions of

Web Science'12, June 22–24, 2012, Evanston, IL, USA.

[1] www.mturk.com
[2] www.taskcn.com
[3] www.crowdflower.com
[4] Our experiments were conducted on Amazon Mechanical Turk, so we adopt their terminology.

different quality; in some problems it may be acceptable to be wrong sometimes, while other problems require solutions of the highest achievable quality. Also, while in some conditions it may be important to obtain solutions instantly, in others a degree of latency may be acceptable.

One major problem in crowdsourcing domains is that the effort level exerted by workers cannot be directly observed by requesters. A worker may attempt to free-ride the system and increase their earnings by lowering the quality and maximizing quantity of their responses [16, 17, 34, 41]. To alleviate this free-riding problem, some researchers have proposed making the workers participate in a contest for the prize [2, 9, 12]. Another possible solution, used by AMT and other crowdsourcing marketplaces, is to employ a reputation mechanis [4, 19, 42]. A requester can *reject* a worker's solution if it falls short of their expectations. In turn, rejection rates can be used to filter out unreliable workers. Effectively, reputation mechanisms allow requesters to choose a desired level of workers' reputation scores and hence expected quality of the work. Reputation mechanisms motivate workers to maintain high reputation in order to access more and better paid tasks.

Another solution to low quality input is aggregation (sometimes called redundancy) [17, 34]. The requester can obtain several responses to each of the HITs, and use this information to improve the quality of the solution. For example, the requester can choose the most popular of the solutions (majority voting) or examine them manually and select the top quality one. In general, the more individual responses there are to each of the HITs the more information is available to construct high quality solutions. However, the costs increase steadily with number of workers assigned to each HIT, while some tasks may be easily solved by a very small number of workers. Therefore it is crucial to decide on the optimal number of workers per HIT.

One major decision the requester must make when designing a crowdsourcing task is the payment offered to the workers. Higher rewards attract more workers and reduce the time required to obtain solutions. However, offering a higher payment does not necessarily increase the quality of the work, as it may encourage the workers to risk their reputation by submitting random or not well thought-out responses. High rewards can also create a psychological pressure decreasing workers' cognitive capabilities.

How can the impact of such factors on the quality of the resulting solutions be determined? And how can the "intelligence" of the crowdsourcing solution be measured and compared against alternatives such as relying on their employees or hiring a consultant? Answering these questions requires a concrete measure of performance, which could be applied both to crowds in a crowdsourcing platform and to single individuals performing the task.

Crowd IQ

We propose that the potential of a crowdsourcing platform to solve mentally demanding problems can be measured using an intelligence test, similarly to how those tests are used to predict individuals' performance in a broad spectrum of contexts, including job and academic performance, creativity, health-related behaviours and social outcomes [14, 18, 24, 35].

We use questions from the widely used IQ test—Raven's Standard Progressive Matrices (SPM) [29, 31]. As our research does not study the IQ of individuals, but rather the IQ of the crowd, the test was not published on AMT as a whole. Instead each of the test questions was turned into a separate HIT. This approach is more typical to the AMT environment, where it is advised[5] to split assignments into the tasks of minimum complexity. The crowd-filled questionnaire is scored using the standard IQ scoring procedure and we refer to it as a *Crowd IQ* - the crowd's potential to solve mentally challenging tasks.

The IQ test we used offers a set of non-trivial problems engaging a range of mental abilities, and thus it measures quality rather than quantity of crowdsourced labour. Using the standard IQ scale we can compare the crowd's performance with the performance of human subjects or other crowds. This measure can also compare the effectiveness of various crowdsourcing settings, such as the reward promised to workers or the reputation threshold for allowing a worker to participate in the task. Also, as the crowd IQ score and an individual's IQ score lie on the same scale, one can compare the performance of a crowd with that of its individual members.

Our Contribution: We use crowd IQ to examine the intellectual potential of crowdsourcing platforms and investigate the factors affecting it. We study the relationship between a crowd's performance and (1) the reward offered to workers, (2) workers' reputation, (3) threatening workers' reputation in case of providing an incorrect response. We show how aggregation of workers' responses improves crowd's performance, and suggest how to further boost it (without increasing the budget) using an adaptive approach assigning more workers to tasks where consensus has not been reached.

Our results provide several practical insights regarding crowdsourcing platforms.

1. Crowdsourcing can lead to higher quality solutions than the work of a single individual.

2. The payment level, task rejection conditions, and workers' reputation all have a huge impact on the achieved performance.

3. Increasing rewards does not necessarily boost performance — moderate rewards lead to highest crowd IQ levels.

[5] http://aws.amazon.com/documentation/mturk/

4. Punishing incorrect responses by decreasing workers' reputation score significantly improves performance.

5. Avoid workers with low reputation - their solutions are usually random.

6. Aggregating workers' opinions can significantly boost performance, especially when using an adaptive approach to dynamically solve ties.

RELATED WORK

Our focus in this paper is on measuring the effect different crowdsourcing settings (e.g. payment, reputation, and aggregation) have on the quality of the obtained results. The impact of the incentive structure in crowdsourcing has been studied in [7, 26] and it was shown that quantity but not quality of work is improved by higher payments. Our results are consistent with those findings.

The recent DARPA red balloon network challenge[6], where the task was to determine the coordinates of a given number of red balloons, has given rise to powerful strategies for recruiting human participants. The MIT team which won the competition used a technique similar to multi-level marketing to recruit participants. The prize money was then distributed up the chain of participants who spotted the balloons.

In other game-oriented crowdsourcing platforms, e.g., games with a purpose [40] (e.g., ESP, Verbosity, Foldit, Google Image Labeler, etc.) the main incentive is recreation and fun, and the tasks to be solved are rather implicit. Yet other platforms have a purely academic flavour and the incentive is scientific in nature. For example, in the Polymath Project [15] the Fields Medallists Terry Tao and Timothy Gowers enable the crowd to provide ideas for deriving proofs to mathematical theorems. Examples such as the DARPA red balloon network challenge or Polymath show that the problems solved by the crowds can be quite advanced and go far beyond micro-tasks. With this in mind, it is pertinent to think about techniques for measuring the corroborated capabilities [8, 11, 20, 21, 32] of a given crowdsourcing system when viewed as a black box of collective intelligence.

Our results are based on decomposing a high-level task (solving an IQ test) into smaller subtasks, such as finding the correct response to each question in the test. Further, we allocate the same subtask to several workers, and aggregate the multiple responses into a single response for that subtask, examining what affects the performance of workers in the subtasks and the performance in the high level task. There is a vast body of literature on aggregating the opinions of multiple agents to arrive at decisions of high quality. Social choice theory investigates joint decision making by selfish agents [36], and game theory can provide insights regarding the impact of the incentive structure on the effort levels of the workers and the quality of the responses.

[6]https://networkchallenge.darpa.mil/Default.aspx

Theoretical analysis from social choice theory, such as Condorcet's Jury Theorem [27] can provide bounds on the number of opinions required to reach the correct response with high probability, and theoretical results from game theory can propose good incentive schemes for crowdsourcing settings [2, 9].The field of judgment aggregation [23] examines how a group of agents can aggregate individual judgments into a collective judgment on interrelated propositions. Collaborative filtering aggregate peoples' opinions to arrive at good recommendations for products or services [5, 13, 22, 33]. Several mechanisms such as Prediction Markets [28] have been proposed for motivating agents to reveal their opinions about the probabilities of future events by buying and selling contracts.

Our methodology relies on similar principles as the above lines of work: we obtain many responses to a standard IQ questionnaire and aggregate them. However, although our principles are similar, our goal was not to examine the properties of the aggregation methods, but rather to determine how to best set up tasks in crowdsourcing environments to achieve high performance. The abovementioned theoretical models make very strong assumptions regarding the knowledge and behaviour of workers. In contrast, we aim to provide practical recommendations for the use of crowdsourcing platforms, so our analysis is empirical.

Our metric is based on the concept of intelligence - a central topic in psychometrics and psychology. There is a strong correlation in people's performance on many cognitive tasks fuelled by a single statistical factor, typically called "general intelligence" [14, 24, 35, 39]. Recent work even extends this notion to "collective intelligence" for the performance of *groups* of people in joint tasks [25, 44]. While our approach relies on similar principles, our performance measure is different from the above mentioned work in that it aggregates multiple opinions of AMT workers, but does not allow them to *interact*. In our setting, AMT workers solve HITs on their own, without discussing the task with the others.

Our performance measure is based on an IQ questionnaire, similarly to [6, 25], which also aggregate individual's responses to an IQ questionnaire (either using majority voting or a machine learning approach). Though our performance measure is similar, our focus is very different. The above work has collected responses in a traditional laboratory environment where individuals have solved an IQ test, attempting to construct tools for evaluating an individual's contribution within the crowd and for optimal team formation. In contrast to that work, we have examined a crowdsourcing setting, trying to determine how crowdsourcing settings such as the incentive structure and task formulation affect the achieved performance. Despite the differences between this paper and the analysis in [6], we note that many of the tools and aggregation methods discussed there, such as the

"contextual IQ" [7] measure for an individual's contribution to the crowd's IQ or the machine learning based aggregation methods, could certainly be applied to a crowdsourcing environment where an individual worker answers several questions.

METHODOLOGY

In our experiments AMT workers were asked to answer questions drawn from an IQ test to establish a proxy for the intellectual potential of the crowd - the crowd IQ score. We also investigated the factors influencing the crowd IQ by modifying the parameters of the crowdsourced tasks—such as the payment, the number of workers per task, worker's reputation, or the rejection rules. We now describe the IQ test and methods used in this study.

Standard Raven Progressive Matrices

The IQ test used in this study, Raven's Standard Progressive Matrices [29, 31] (SPM), was originally developed by John C. Raven [30] in 1936. SPM is a non-verbal multiple-choice intelligence test based on Spearman's theory of general ability [39]. Raven's SPM and its other forms (e.g., Advanced and Colored Progressive Matrices) are among the most commonly used intelligence tests, in both research and clinical settings [31].

SPM consists of 60 questions, each of which contains a square array of patterns called "matrix" with one element missing and a selection of 8 possible responses. Matrices are separated into five sets (called A,B,C,D,E) of 12 questions each, where within each set the questions are arranged in increasing difficulty.

The sets themselves are arranged in order of increasing difficulty, with an overlap in difficulty levels. For example although the questions in set B are generally more difficult than those in set A, the last items in set A are more difficult than first items in set B. In this study we left out the two easiest sets of matrices (A and B) following the standard procedure (*starting rule*) employed when administering the test to individuals with average or above average IQ.

Data collection and workers

Each of the 36 IQ test questions was turned into a separate HIT (Human Intelligence Task in AMT terminology) which required workers to choose a correct response. Those HITs were published on AMT in January and February 2011 under several experimental conditions. In each of the experimental conditions we requested five solutions per HIT. We limited access to HITs of this study to workers from the US that had previously submitted at least 200 HITs on AMT. There were 175 unique workers in our study, and each of them submitted 4.1 HITs on average. To prevent workers from participating in more than one testing condition of our experiment, the HITs belonging to only one condition were available at any given time. Also, we removed completed HITs from the dataset that were submitted by workers who appeared in more than one of the experimental conditions. To minimize the biases related to the weekly or daily fluctuations in AMT efficiency, experimental conditions were published on different dates but on the same weekday and at the same time of day.

Note that AMT workers can see a HIT before deciding to work on it and can also decline to respond (*return HIT*) without any consequences for their reputation. This has a potential to boost AMT's IQ, comparing with the traditionally administered questionnaires where respondents have to answer all of the questions. However, this feature is inherent to the AMT environment and, as such, was desirable in this study.

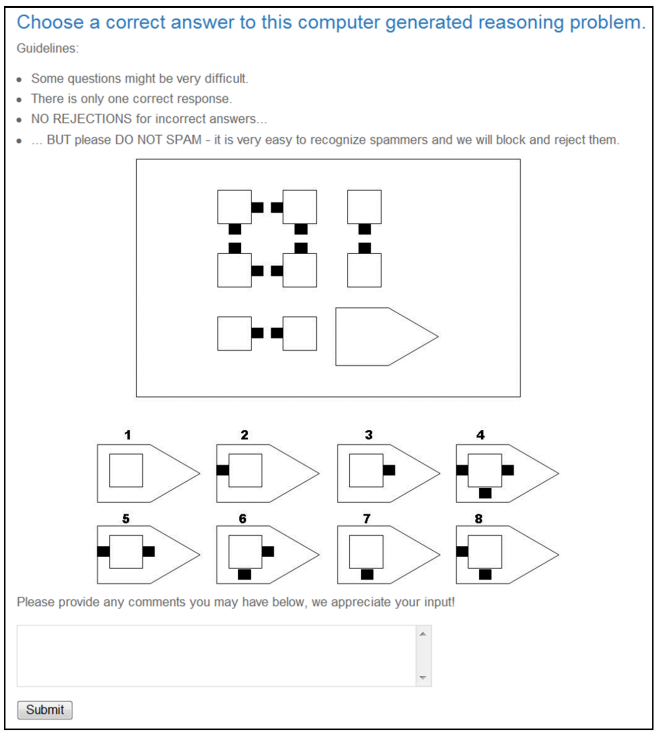

Figure 1. Sample Human Intelligence Task as used in this study. Note that as the SPM items are copyright protected, the IQ question presented here is not an actual SPM item, but a similar one.

Design of the HIT

We used two distinct payment models and thus needed two slightly different HIT designs. In the first payment model, we offered payment only for correct responses but did not reject HITs in the case of an incorrect response. Effectively, the reputation of the workers was not affected by incorrect responses. An example of such HIT is presented on Figure 1; note that it states that there are "*NO REJECTIONS for incorrect answers*". In the second payment model, HIT's description stated that incorrect responses will be rejected affecting the worker's

[7] Contextual IQ attempts to measure an individual's contribution to the crowd's IQ using the Shapley value concept from game theory, and employs a power index approximation algorithm [3, 37, 38].

reputation ("*Incorrect answers will be REJECTED*"). In fact, no workers were hurt during this experiment as all of the HITs were accepted after we collected all of our data. In both conditions we have attempted to limit the number of dishonest submissions (spamming) by stating that "*.. it is very easy to recognize spammers and we will block and reject them*". To avoid revealing the nature of this study, the HITs were described as "*computer generated reasoning problems*".

Note that we used the first payment model and HIT design in all but one experiment (focused on the effect of rejection risk) in order to reduce the potential stress imposed on the workers participating in this study. The risk of rejection and resulting decrease in reputation acts as a strong deterrent against free-riding the system by submitting responses of poor quality. However, in the case of the intellectually demanding questions, even those investing a lot of time and effort may get the wrong answer and thus may experience a certain degree of anxiety. We show that the threat of rejection has a large positive effect on the crowd IQ, but we avoided imposing it whenever possible.

Scoring

We requested five solutions to each of the 36 IQ questions (one from five different workers) across all of the experimental conditions. The sum of correct solutions was divided by five, which is equivalent to computing an average of scores on each of the five complete solutions. The scores were compared between conditions using a Wilcoxon signed-rank test. The analysis was performed on the level of individual IQ questions by comparing the number of correct responses between the conditions.

Crowdsourced solutions were scored using a standard scoring procedure described in the SPM test manual [29]. The manual provides lookup tables which allows translating the number of correct responses (raw score) into an IQ score. The IQ scale characteristic for SPM, similarly to most other intelligence tests, is standardized on a representative population to follow a normal distribution with an average score of 100 and standard deviation of 15. As we did not control for the age of the workers, we used the most demanding norms—those for the 16 and 17 year old subjects. The norms of all the other age groups are lower which would result in higher IQ scores for any given raw score. We refer to the obtained score as the *crowd's IQ*[8].

RESULTS

We now present the results of our study. We show how the crowd IQ is affected by the settings of crowdsourcing tasks discussed before, including (1) the threat of rejection, (2) reward level, and (3) the reputation of the workers. Then we investigate how aggregating multiple responses to each HIT improves the crowd IQ and demonstrate the boost in performance introduced by an adaptive sourcing scheme.

Rejection Risk

We examine the influence of the rejection threat on the crowd IQ by comparing the two different payment conditions discussed in the Design of the HIT section. Both offered a reward of $0.10 for the correct responses only and were accessible to workers with reputation of at least 98%. In the rejection condition, we stated that the incorrect responses will be rejected and thus would adversely affect the worker's reputation. The crowd IQ scores were calculated using the method described in the Scoring section.

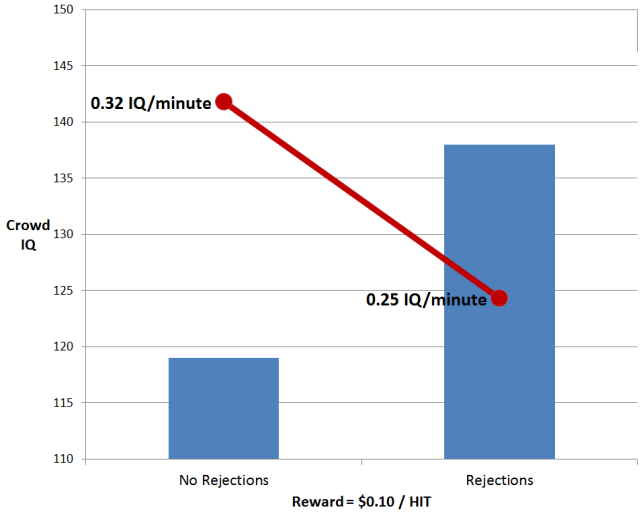

Figure 2. Crowd IQ and crowd IQ per minute in the "rejection" and "no rejection" approaches (one worker per HIT, min. reputation 98%). Difference in crowd IQ is significant at the p<0.001 level; $Z = 3.62$.

Figure 2 shows the effect of the rejection rule on the crowd IQ. The AMT crowd IQ is far higher (by 20 IQ points) under the presence of the rejection threat. The magnitude of this difference is best appreciated by comparison with the frequencies of such scores in the general population. Whereas the crowd IQ score achieved in the no rejection condition (119 IQ) is exceeded by roughly one in ten people in the general population, the IQ score achieved under the risk of rejection (138 IQ) is exceeded only by one in two hundred people in the general population.

However, as workers are concerned with their reputations, the threat of rejection also reduces the number of workers willing to perform the task, making the crowd slower in generating solutions. While the rejection rule was present, AMT needed nearly 50% more time to finish the task (9:05 hours versus 6:12 hours in the no-rejection condition). Consequently the number of correct solutions per minute, as expressed by the crowd IQ score

[8]We use the same terminology as [6], referring to the score of the aggregated responses of a crowd as the "crowd's IQ". Obviously, when comparing the scores to the norms of an individual (sampled from the general population), the crowd's IQ score has quite different properties, as discussed in [6].

divided by time needed to complete the task, is higher for the no-rejection condition.

Payment levels

We now examine the relation between the offered payment and the performance of the crowd. We used the no-rejection rule discussed in the Design of the HIT section, four different levels of payment, and limited access to the task to workers with reputation of at least 98%. The crowd IQ scores were calculated using a method described in the section on scoring. Figure 3 shows the change in the crowd IQ and IQ points harnessed per minute across the four payment levels—0.01$, 0.05$, 0.10$, and 0.20$.

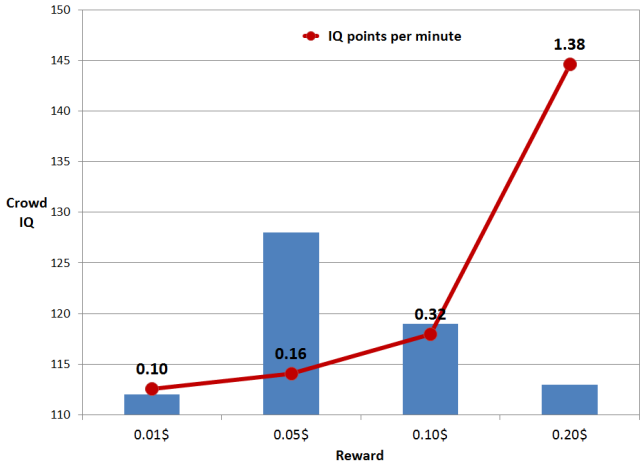

Figure 3. The crowd IQ, and crowd IQ per minute as functions of payment levels (one worker per HIT, 0.10$ per HIT, min. reputation 98%, no-rejection). Difference in the crowd IQ between 0.01$ and 0.20$ levels is not significant at the p<0.05 level. Other differences in the crowd IQ are significant at the p<0.05 level.

Increasing the payment from 0.01$ to 0.05$ boosts the crowd IQ. However, further increases to 0.10$ or 0.20$ per HIT leads to a decrease in the crowd IQ. The time needed to obtain the full solution to the task decreases rather sharply with amount paid per HIT - from more than 18 hours in 0.01$ condition to mere 15 minutes in the 0.20$ condition. Effectively, paying 0.20$ per HIT yields around 14 times more IQ points per minute than paying 0.01$. However, note that the total crowd IQ is comparable between those two payment levels while cost increases by a factor of twenty. If speed is not a priority it seems advantageous to pay less and wait longer to obtain a solution.

The results discussed above suggest that crowd IQ does not increase steadily with the payment offered. One possible cause is that more lucrative tasks attract a greater number of free-riders submitting random or not well thought over responses hoping for luck or sloppy quality control. Additionally, high payment may increase psychological pressure on the workers thus affecting their mental performance.

Using reputable workers

We now examine the relation between the crowd IQ and the reputation of workers. The reputation of a worker is captured by the acceptance rate of their HITs (i.e. one minus the proportion of their HITs that were rejected by the requesters). We examine the crowd IQ obtained under different reputation thresholds, i.e., when we only allowed workers with an acceptance rate beyond a certain level to participate in the task.

The crowd IQ scores were calculated using a method described in the Scoring section. Figure 4 shows the crowd IQ achieved with payment of .05$, under the no-rejection rule and using a single worker per HIT across five ranges of acceptance rates. Using individuals with a high reputation score leads to much higher crowd IQ, indicating that using workers with acceptance rates lower than 95% makes little sense, given that more reputable workers are willing to engage in the task.

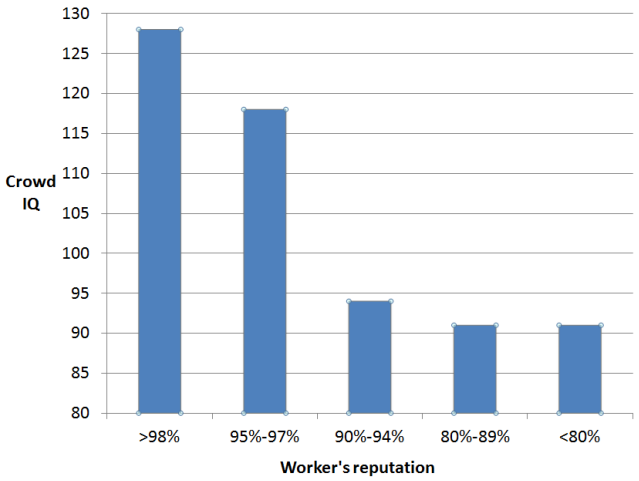

Figure 4. The crowd IQ as a function of workers' reputation scores (one worker per HIT, 0.05$, no-rejection). Differences among the three lowest reputation ranges are statistically insignificant. All other differences are statistically significant at the p<0.001 level.

Aggregating multiple responses

The results presented in the previous sections were based directly on the number of correct responses submitted by the workers, divided by number of workers per HIT—which is equivalent to requesting AMT to provide a single solution per HIT. However, we now demonstrate that it is advisable, resources permitting, to request more solutions to each of the HITs and to aggregate them to increase the crowd IQ.

Figure 5 shows the relation between the number of workers per HIT and the resulting crowd IQ. It depicts significant gains from the first few workers and diminishing returns from each additional worker. To create Figure 5 we have used a simple majority voting approach, that chose the most common response from up to 24 individual responses to each of the HITs (similarly to [6, 25]). If two or more responses are selected by the workers an

equal number of times (a tie), one of those was chosen at random.

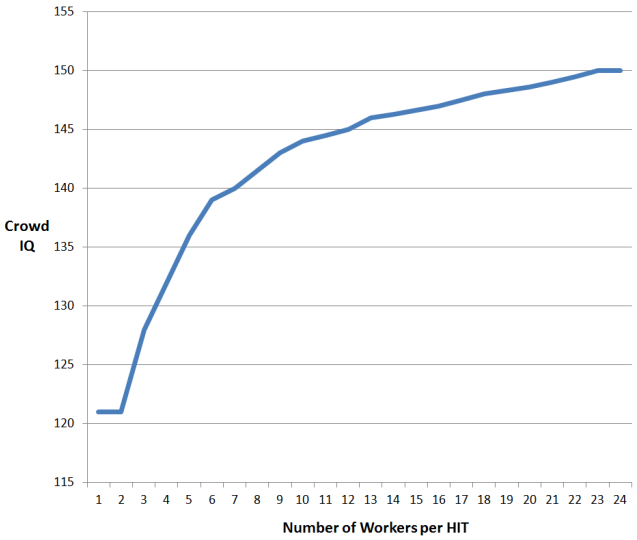

Figure 5. The crowd IQ as a function of the number of workers per HIT (all payment conditions included, reputation > 98%). For each number of workers the plot shows the average crowd IQ score over 1,000 sub-samples of responses.

To smooth the line and decrease the selection bias, our procedure was repeated 1000 times for randomly selected response subsamples of any given size and the resulting IQ scores were averaged. Responses used in this experiment come from all of the experimental conditions based on the workers with more than 98% HITs accepted, as described before.

AMT using a single worker per HIT can achieve a higher IQ score than 91% of individuals in the general population (122 IQ). With more workers available the advantage of the crowd grows rather quickly. Even more strikingly, only one person in a thousand in the general population has an IQ higher than the AMT using 12 workers per HIT (145 IQ).

Figure 6 was derived from Figure 5 and shows the average increase in the crowd IQ with each additional worker. Using two workers per HIT does not increase the crowd IQ because in case of a disagreement two workers would always produce a tie that has to be solved by choosing a random of two responses. Up to the regime of six workers per HIT, each additional worker per HIT adds more than 3.5 crowd IQ points, but gains decrease with the increasing size of the crowd.

Solving ties dynamically

In the previous section we showed that the crowd IQ increases with the number of workers per HIT. However, for some HITs, one of the responses (not necessarily the correct one) may be clearly preferred by the workers and thus increasing the number of votes is unlikely to change crowd's decision. Consequently, instead of increasing the

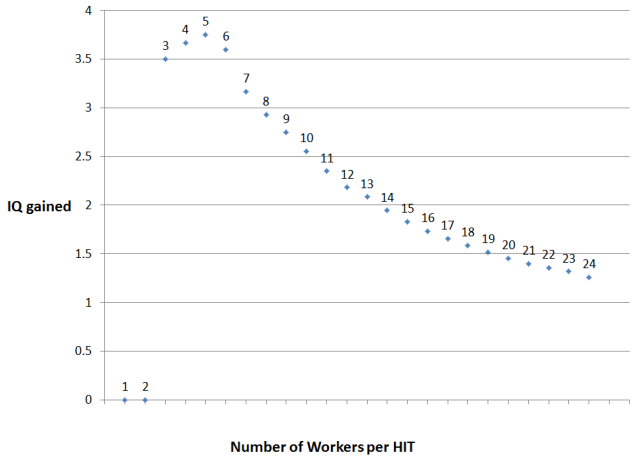

Figure 6. Marginal increase in the crowd IQ with each additional worker, derived from Figure 5

number of workers per HITs in all of the HITs (*fixed approach*), one could assign additional workers only to those HITs where consensus has not been reached.

Based on this notion, we propose a simple *adaptive approach* where a fixed number of workers is assigned to each of the tasks (*initial workers*) and their votes are recorded. In case of a tie, additional workers are assigned, until the consensus has been reached. As number of necessary additional workers varies between the questions, the average number of responses per HIT may be a fraction rather than integer.

Figure 7 compares the performance of the fixed and adaptive approaches. The crowd IQ values (on the Y axis) were calculated using both the adaptive and fixed approaches. The X axis shows the average number of workers per HIT, while the labels above the points, relating to the adaptive approach signify the initial number of workers (between two and ten).

The adaptive approach has a clear advantage over the fixed one, especially for small numbers of initial workers when ties are more likely to occur. For example, by assigning two workers to each HIT and dynamically solving ties by adding additional workers, one can achieve a crowd IQ score of 136 at the expense of 2.9 workers per HIT. This IQ score is comparable with the one achieved in the fixed approach using five workers per HIT and it is eight IQ points higher than the fixed approach with three workers per HIT.

CONCLUSIONS AND LIMITATIONS

We examined combining multiple responses to questions from a standard IQ test into a single filled questionnaire, and scoring this solution using the standard scoring key for IQ tests. We refer to this score as the crowd IQ, and proposed using it as a measure of the quality of solutions obtained from crowdsourcing platforms. We quantitatively analysed several factors affecting the quality of the obtained solution, using the crowd IQ as our metric.

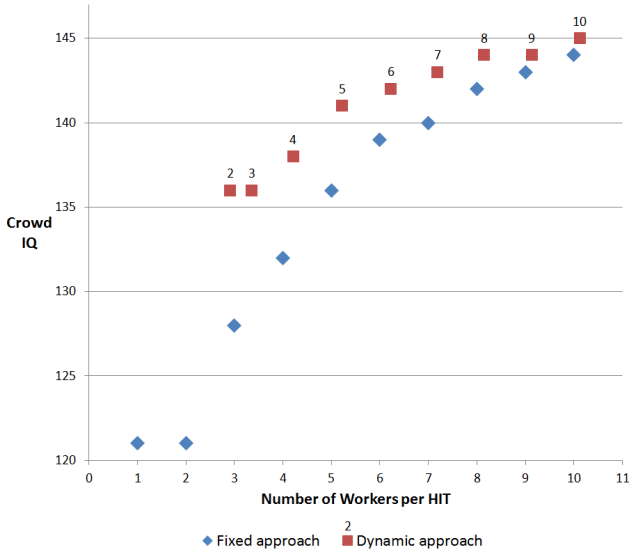

Figure 7. The crowd IQ and number of workers per HIT while using affixed and adaptive crowdsourcing. The fixed approach is equivalent to Figure 5. The numbers above the squares represent the number of initial workers.

Our results indicate that punishing wrong responses with a rejection and thus decreasing workers' reputation score, significantly increases the crowd IQ. However, this comes at the cost of attracting fewer workers, and thus increasing the amount of time required to solve the entire task. Additionally, our results show that the effect of the reward on crowd IQ is not monotone. Both too high and too low rewards significantly affect the performance, potentially by encouraging the free-riding behaviour and increasing the psychological pressure on the workers. However, as higher rewards attract more workers, the time that AMT requires to solve the problem decreases quickly with payment offered. Further, tightening the reputation requirements achieves solutions of higher quality. Finally, using the adaptive sourcing schemes allows utilizing the available budget more effectively. By focusing on questions where consensus has not yet been reached, the requester can achieve extremely high performance even for relatively small crowds.

Recommendations for effective crowdsourcing: Our research provides several useful recommendations for effective crowdsourcing. Crowdsourcing solutions appear to have a great potential, as even small crowds outperform most individuals from the general population when the task is designed properly. Below, we provide several recommendations for successful crowdsourcing stemming from this research.

First, experiment and monitor the performance. Our results suggest that relatively small changes to the parameters of the task may result in great changes in crowd performance. Changing parameters of the task (e.g. reward, time limits, reputation rage) and observing changes in performance may allow you to greatly increase performance.

Second, make sure to threaten workers' reputation by emphasizing that their solutions will be monitored and wrong responses rejected. Obviously, in a real-world setting it may be hard to detect free-riders without using a "gold-set" of test questions to which the requester already knows the correct response. However, designing and communicating HIT rejection conditions can discourage free riding or make it risky and more difficult. For instance, in the case of translation tasks requesters should define what is not acceptable (e.g. using Google Translate) and may suggest that the response quality would be monitored and solutions of low quality would be rejected.

Third, do not over-pay. Although the reward structure obviously depends on the task at hand and the expected amount of effort required to solve it, our results suggest that pricing affects not only the ability to source enough workers to perform the task but also the *quality* of the obtained results. Higher rewards are likely to encourage a free-riding behaviour and may affect the cognitive abilities of workers by increasing psychological pressure. Thus, for long term projects or tasks that are run repeatedly in a production environment, we believe it is worthwhile to experiment with the reward scheme in order to find an optimum reward level.

Fourth, aggregate multiple solutions to each HIT, preferably using an adaptive sourcing scheme. Even the simplest aggregation method - majority voting - has a potential to greatly improve the quality of the solution. In the context of more complicated tasks, e.g. translations, requesters may consider a two-stage design in which they first request several solutions, and then use another batch of workers to vote for the best one. Additionally, requesters may consider inspecting the responses provided by individuals that often disagree with the crowd - they might be coveted geniuses or free-riders deserving rejection.

Limitations: Our approach has several limitations. First, the performance of a crowd in solving an IQ test may not be predictive about how the same crowd would perform in other tasks. While the correlation between an individual's performance in an IQ test and their performance in many other settings has been shown to be strong in many studies [14, 18, 24, 35], very few such studies have been conducted for crowds [25, 44].

Next, in our setting the workers have expressed their opinions independently. While this is typical for many crowdsourcing environments, our results may not hold in domains where members interact while solving the task. For example, certain collaborative systems such as Wikipedia foster active collaboration between participants, where a single solution is updated many times by various individuals, who comment on the work of their peers. Thus, further research is required to examine the

factors that affect performance in such interactive and potentially creative environments.

Yet another limitation of our approach is the fact that our study was conducted on the AMT crowdsourcing platform, rather than in a controlled lab environment. A lot of the effects we noticed can be explained by self-selection biases or free-riding behaviours that are common in such platforms, but may not hold in other environments. Conducting a similar study in a lab setting may allow controlling for such factors and focusing on specific effects. Also, while it is possible to apply the same methodology to other crowdsourcing platforms, it is not guaranteed that the same results would be obtained.

Finally, our aggregation approach relies on simple majority voting. Better performance may be achieved when using a more sophisticated aggregation. For example, more weight could be given to individuals of a higher reputation, or participants' confidence in their response could be used as another factor.

Future Research: Many fascinating questions are left open for further investigation. Could similar results hold for tasks other than answering IQ tests? What are the best methods for aggregating individual solutions? Can further information regarding a worker be used to predict their contribution in such a crowdsourcing setting? Would allowing interactions between group members improve or hinder the performance? Which interaction mechanisms are suitable for such settings? How should the incentives in such an interactive environment be structured to optimize performance? We hope that our use of the crowd IQ to measure performance may encourage the research community in the area of collective intelligence to consider making use of the psychometric tools to better understand what collective intelligence can do.

REFERENCES

1. L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.
2. N. Archak and A. Sundararajan. Optimal design of crowdsourcing contests. *ICIS 2009 Proceedings*, 200, 2009.
3. Y. Bachrach, E. Markakis, E. Resnick, A.D. Procaccia, J.S. Rosenschein, and A. Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multiagent Systems*, 2010.
4. Y. Bachrach, A. Parnes, A.D. Procaccia, and J.S. Rosenschein. Gossip-based aggregation of trust in decentralized reputation systems. *Autonomous Agents and Multi-Agent Systems*, 19(2):153–172, 2009.
5. Y. Bachrach, E. Porat, and J.S. Rosenschein. Sketching techniques for collaborative filtering. *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 2016–2021, 2009.
6. Y. Bachrach, Graepel T., Kasneci G., Kosinski M., and J. Van-Gael. Crowd IQ - aggregating opinions to boost performance. In *AAMAS*, 2012.
7. W.A.M. Borst. *Understanding Crowdsourcing: Effects of motivation and rewards on participation and performance in voluntary online activities*. PhD thesis, Erasmus University Rotterdam, 2010.
8. A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.
9. D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 119–128. ACM, 2009.
10. A. Doan, R. Ramakrishnan, and A.Y. Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.
11. A. Galland, S. Abiteboul, A. Marian, and A. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.
12. A. Gao, Y. Bachrach, Key P., and T. Graepel. Quality expectation-variance tradeoffs in crowdsourcing contests. *AAAI*, 2012.
13. D. Goldberg, D. Nichols, B.M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.
14. L.S. Gottfredson. Why g matters: The complexity of everyday life. *Intelligence*, 24(1):79–132, 1997.
15. T. Gowers and M. Nielsen. Massively collaborative mathematics: The Polymath project. *Nature*, 461(7266):879–881, October 2009.
16. P.Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35. Association for Computational Linguistics, 2009.
17. P.G. Ipeirotis, F. Provost, and J. Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.
18. A.R. Jensen. The g factor: The science of mental ability. *London: Westport*, 1998.
19. A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007.

20. G. Kasneci, J. Van Gael, D. H. Stern, and T. Graepel. Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In *WSDM*, pages 465–474, 2011.

21. G. Kasneci, J. Van Gael, D. H. Stern, R. Herbrich, and T. Graepel. Bayesian knowledge corroboration with logical rules and user feedback. In *ECML/PKDD (2)*, pages 1–18, 2010.

22. G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

23. C. List and C. Puppe. Judgment aggregation: A survey. *Handbook of Rational and Social Choice*, 2009.

24. D. Lubinski. Introduction to the special section on cognitive abilities: 100 years after spearman's (1904) "general intelligence, objectively determined and measured". *Journal of Personality and Social Psychology*, 86(1):96, 2004.

25. J.A. Lyle. Collective problem solving: Are the many smarter than the few? 2008.

26. W. Mason and D.J. Watts. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter*, 11(2):100–108, 2010.

27. A. McLennan. Consequences of the condorcet jury theorem for beneficial information aggregation by rational agents. *American Political Science Review*, pages 413–418, 1998.

28. D.M. Pennock and R. Sami. Computational aspects of prediction markets, 2007.

29. J.C. Raven. Standard progressive matrices plus, sets a-e.

30. J.C. Raven. *Progressive matrices*. Éditions Scientifiques et Psychotechniques, 1938.

31. J.C. Raven. The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1):1–48, 2000.

32. V.C. Raykar, S. Yu, L.H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

33. B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

34. G. Sautter and K. Böhm. High-throughput crowdsourcing mechanisms for complex tasks. *Social Informatics*, pages 240–254, 2011.

35. F.L. Schmidt and J. Hunter. General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86(1):162, 2004.

36. A. Sen. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181, 1986.

37. L. S. Shapley. A value for n-person games. *Contrib. to the Theory of Games*, pages 31–40, 1953.

38. L. S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48:787–792, 1954.

39. C. Spearman. The abilities of man. 1927.

40. L. von Ahn. Games with a Purpose. *Computer*, 39(6):92–94, June 2006.

41. M. Vukovic. Crowdsourcing for enterprises. In *Services-I, 2009 World Conference on*, pages 686–692. Ieee, 2009.

42. A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proc. 7th Int. Workshop on Trust in Agent Societies*, 2004.

43. J. Wolfers and E. Zitzewitz. Prediction markets. Technical report, National Bureau of Economic Research, 2004.

44. A.W. Woolley, C.F. Chabris, A. Pentland, N. Hashmi, and T.W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686, 2010.