Realizing Impact Sourcing by Adaptive Gold Questions: A Socially Responsible Measure for Workers' Trustworthiness

Kinda El Maarry[1], Ulrich Güntzer[2], Wolf-Tilo Balke[1]

[1] IFIS, TU Braunschweig, Braunschweig, Germany
{elmaarry, balke}@ifis.cs.tu-bs.de
[2] Inst. f. Informatik, Universität Tübingen, Tübingen, Germany
ulrich.guentzer@informatik.uni-tuebingen.de

Abstract.
In recent years, crowd sourcing has emerged as a good solution for digitizing voluminous tasks. What's more, it offers a *social* solution promising to extend economic opportunities to low-income countries, alleviating the welfare of poor, honest and yet uneducated labor. On the other hand, crowd sourcing's virtual nature and anonymity encourages fraudulent workers to misuse the service for quick and easy monetary gain. This in turn compromises the quality of results, and forces task providers to employ strict control measures like gold questions or majority voting, which may gravely misjudge honest workers with lower skills, ultimately discarding them from the labor pool. Thus, the problem of fairly distinguishing between fraudulent and honest workers lacking educational skills becomes vital for supporting the vision of Impact Sourcing and its social responsibility. We develop a technique with socially responsible gold questions as an objective measure of workers' trustworthiness, rather than a mere discarding mechanism. Our statistical model aligns workers' skill levels and questions' difficulty levels, which then allows adapting the gold questions' difficulty for a fair judgment. Moreover, we illustrate how low-skilled workers' initial payloads, which are usually discarded along with the worker, can be partially recovered for an increased economic gain, and show how low-skilled workers can be seamlessly integrated into high-performing teams. Our experiments prove that about 75% of misjudged workers can be correctly identified and effectively be integrated into teams with high overall result correctness between 70-95%.

Keywords: crowd sourcing, impact sourcing, fraud detection, quality control

1 Introduction

Crowd sourcing platforms can distribute cognitive tasks requiring human intelligence through digital gateways, which can flexibly tap into huge international workforces. In a nutshell, it creates a win-win opportunity where task providers can cut down their costs through cheaper services, while simultaneously providing economic opportunities to hired workers. Coupled with *Impact Sourcing* it could play a key role in advancing the economic development of low-income countries, alleviating the welfare of less fortunate individuals, as well as connecting them to the global economy. Impact Sourcing, the socially responsible arm of the information technology outsourcing industry [1], specifically aims at employing people at the bottom of the pyramid, who are disadvantaged on an economical, educational and accordingly skill-wise level. However, the highly distributed nature, virtual and anonymous setup of crowd sourcing platforms, along with the short term task contracts they offer open doors for *fraudulent workers*, who can simply submit randomly guessed answers, in hope of going undetected. The inclusion of such workers of course jeopardies the overall credibility of the returned quality. And with manual checking being both costly and time consuming, this directly invalidates the main gains of crowd sourcing. Hence, task providers are forced to employ strict control measures to exclude such workers, ensure high quality results, and get good return on their investment. However, these measures befall honest, yet low-skilled workers, too. In fact, anecdotal evidence from our own previous work [2] shows that by completely excluding workers from two offending countries, where a high number of fraudulent workers were detected, the overall result correctness instantly saw a 20% increase. Needless to say, this simultaneously excluded many honest workers in those two countries as well.

Indeed, the positive social impact of the Impact Sourcing model is immense, where almost half of the world's population lives on less than $2.50 a day, and 1.8 billion people can't access a formal job[1]. But also with Impact sourcing this huge task force may ultimately fall into a vicious cycle: even with simple task training mechanisms offered by platforms like CrowdFlower, the opportunity provided by crowd sourcing is biased by quality control measures towards educated workers. In fact, quality measures tend to repeatedly exclude uneducated, low-skilled workers. Not giving them a chance at improving their skills leaves them prey for constant exclusion.

Common currently deployed *quality control measures* include gold questions, majority votes, and reputation based systems. Of course, all such control measures are susceptible to the ultimate downside of misjudging honest low-skilled workers. Accordingly, in this paper we develop an objective socially responsible measure of workers' trustworthiness: *adaptive gold questions*. Basically, an initial set of balanced gold questions (i.e. covering all difficulty levels) is used as a mechanism for determining the skill level of a worker rather than a discarding mechanism. Next, a second round of adapted gold questions, whose difficulty levels are within the estimated skill level of the corresponding worker, are injected. The underlying assumption is that, although low-skilled workers may fail the correctness threshold set for the balanced gold questions, since they surpass their own skill level, they should succeed at gold questions, which have been adapted to their lower skill levels. On the other hand, fraudulent workers would also fail such adaptive gold questions, since their responses to both sets of balanced and adaptive gold questions will be random.

To adapt gold questions, our method requires two parameters: workers' skill levels and difficulties of questions. To that end, we make use of psychometric item response theory (IRT) models: in particular, the Rasch Model for estimating these parameters. Our experiments show that around 75% honest misjudged workers can be correctly identified and the payloads that would have been discarded with the worker can be

[1] http://www.impacthub.org/

partially recovered i.e. tasks in the payload within a low-skilled worker's ability. Furthermore, we investigate heuristics for forming high-performing skill-based teams, into which low-skilled workers can be later integrated to ensure high quality output.

2 Related Work

The social model of Impact Sourcing was first implemented by *Digital Divide Data (DDD)*[2] back in 2001, and ever since has been adopted by many crowd sourcing platforms such as *Samasource*[3], *RuralShores*[4], or *ImpactHub*[1]. Crowd sourcing provides an accessible solution to both: companies having digital intelligent problems (e.g. web resource tagging [3], completing missing data [4], sentiment analysis [5], text translation [6], information extraction [7], etc.) and underprivileged honest workers lacking high skills. But for actually profiting from this win-win situation, the challenge of identifying fraudulent workers and their compromising contributions must be met.

A rich body of research addresses the quality problem in crowdsourcing. Currently employed solutions include aggregation methods, which rely on redundancy as means to improving the overall quality: By assigning the same task to several workers, the correct answer can be identified through aggregation, e.g. majority voting. Nevertheless, this has been shown to have severe limitations, see e.g. [8]. This was followed by Dawid and Skene [9], who applied an expectation maximization algorithm to consider the responses' quality based on the individual workers. Focusing on such error rates, other approaches emerged such as: a Bayesian version of the expectation maximization algorithm approach [10], a probabilistic approach taking into account both the worker's skill and the difficulty of the task at hand [11], or an even more elaborate algorithm trying to separate unrecoverable error rates from recoverable bias [12].

Another class of solutions focuses on eliminating unethical workers throughout longer time scales. This can be achieved through constantly measuring workers' performance via a reputation-based system (based on a reputation model [13-14], on feedback and overall satisfaction [14], or on deterministic approaches [15], etc.) or through injecting gold questions in the tasks. Except, reliably computing the workers' reputation poses a real challenge, and as we will show in section 3, both techniques are susceptible to the ultimate downside of misjudging honest low-skilled workers. In contrast, we apply gold questions as a socially responsible measure of workers' trustworthiness to measure their skill level rather than as a discarding mechanism.

Furthermore, monetary incentives as means of quality control have been investigated. But the implementation of such an approach proves to be tricky, where low paid jobs yield sloppy work, while high paid jobs attract unethical workers [16].

It is important to note how tightly coupled our work is with the IRT Paradigm [17] in psychometrics, which enables us to focus on the workers' capabilities. We employ the Rasch model [18] to estimate the tasks' difficulty and workers' skill. This allows us to address the principal concern of Impact Sourcing: distinguishing honest low-

[2] http://www.digitaldividedata.com/about/

[3] http://www.samasource.org/

[4] http://ruralshores.com/about.html

skilled workers from unethical workers. Perhaps most similar to our work is the model presented in [11], which is also based on the IRT paradigm: GLAD – a generative model of labels, abilities and difficulties – iteratively estimates the maximum likelihood of the worker's skill, question's difficulty, as well as the worker's correctness probability computed by EM (Expectation-Maximization approach). GLAD's robustness wavers though when faced with unethical workers, especially when they constitute more than 30% of the task force [19]. In contrast, we focus on detecting the workers' skill level to adapt future gold questions to be injected, which then enables us to identify with sufficient robustness honest workers who are merely low-skilled.

Other research focusing on estimating one or more parameters include, Dawid and Skene [9], who considered the worker's skill and utilized confusion matrices as an improved redundancy technique form. The downfall as pointed out and addressed by Ipeirotis [20] is the underestimation of the workers' quality, who consistently give incorrect results. Considering two parameters, both the workers' abilities as well as the inference of correct answers was investigated in [21]. Except, the difficulty of the task at hand, which in turn influences the workers' perceived skill level is neglected.

3 Motivational Crowd sourcing in a Laboratory-based Study

For self-containment, we briefly detail in this section one of our earlier experiments in [22]. To acquire ground truth, we set up a small-scale laboratory experiment, with a total of 18 volunteers. In this paper we formulate our Human Intelligent Tasks (HITs) over an American standardized test for college admission: the Graduate Record Examination (GRE) crawled dataset (http://gre.graduateshotline.com), namely the verbal practice questions section. The task then is to select the correct definition out of 4 choices for a given word. Given a set of 20 multiple choice questions, volunteers were asked to answer questions twice. In the first round, they should just randomly select answers while in the second round, they should consider the questions and answer them to the best of their knowledge. Accordingly, the dataset can be divided into honest and unethical workers.

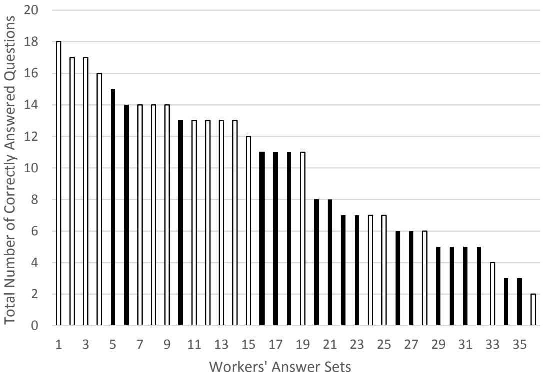


Fig.1. Truthful versus random responses

Figure 1 sorts all workers' answers according to the respective total number of correct answers achieved over 20 questions. Although no worker got all 20 answers correct, it comes as no surprise that truthful answers (58.6%) tend to be more correct than random answers (40%). Furthermore, even though the dataset is in no way biased, random responses at times produced better overall results. Consider the top 10 workers getting the most correct answers in figure 1. In a reputation based system, the worker at rank 5 (scoring 15 correct answers) would be given a higher reputation score than workers on ranks 6 to 9 (scoring 14 correct answers). Yet here, 3 workers at least tried to answer correctly.

Furthermore, with the common 70% correctness threshold set, gold questions would eliminate 61% honest workers (i.e. 11 workers) and 88% of the unethical workers (i.e. 16 workers). Though gold questions are more biased to penalize unethical workers, still the bias is small, and a significant number of honest workers are penalized too.

4 Identifying Low-Skilled Workers

As shown, gold questions tend to misjudge honest low-skilled workers and can be bypassed by unethical workers. In this section, we provide a short overview of the underlying statistical *Rasch Model* (RM), which is used to align workers' skill levels and questions' difficulty levels to adapt the gold questions' difficulty for a fairer judgment and a socially responsible measure that can identify low-skilled workers.

4.1 The Rasch Model

The intrinsic nature of crowdsourcing involves many human factors. This in turn directed our attention to psychometrics – the science assessing individual's capabilities, aptitudes and intelligence – and it's IRT classes, namely, the Rasch model (RM). Basically, RM computes the probability P_{ij} that the response of a worker $\omega_i \in W$ to a given task $t_j \in \mathbb{P}$ is correct as a function of both: 1) his/her ability θ_{ω_i}, and 2) the difficulty of the task β_{t_j}. Assuming a binary setting, where a worker's response $x_{ij} \in \{0,1\}$ is known (where 0 indicates an incorrect response and 1 a correct response), RM's dichotomous case can be applied. Simply put, both the RM's parameters: a worker's ability θ and a task's difficulty β are depicted as latent variables, whose difference yields the correctness probability *P*.

Definition 1: (Rasch Model for Dichotomous Items) given a set of workers W= $\{\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, \dots, \boldsymbol{\omega_n}\}$, where $|\boldsymbol{W}| = \boldsymbol{n}$ and a HIT $\mathbb{P} = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \dots, \boldsymbol{t_m}\}$, where $|\mathbb{P}| = \boldsymbol{m}$. Assume $\boldsymbol{\omega_i} \in \boldsymbol{W}$ and $\boldsymbol{t_j} \in \mathbb{P}$, then the correctness Probability $\boldsymbol{P_{ij}}$ can be given as follows:

$$P_{ij} = P_{ij}(x_{ij} = 1) = \frac{exp(\theta_{\omega_i} - \beta_{t_j})}{1 + exp(\ \theta_{\omega_i} - \beta_{t_j})}$$

This can also be reformulated, such that the distance between $\theta_{\omega_i}$and $\beta_{t_j}$is given by the logarithm of the odds ratio, also known as the log odd unit *logit*.

$$log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \theta_{\omega_i} - \beta_{t_j}$$

The difficulty of a question with a logit value of 0 is average, where a negative logit value implies an easy β_{t_j} and a low $\theta_{\omega_i}$and vice versa. Accordingly, the correctness probability of a worker's response is high when his ability exceeds the corresponding task's difficulty. A special vital feature of RM is its emphasis on the "objective" measurement of (θ, β) [23]. That is, measurement of both θ and β should be independent respectively of $\mathbb{P}$ and W.

4.2 Adapting the Gold Questions based on RM's alignment

Initially, a balanced set of gold questions $\mathbb{G}^B$are injected in an initial payload $\mathbb{P}_a$ to which worker ω is assigned to. However, failing the correctness threshold $\mathbb{C} > 70\%$ (i.e. worker fails on more than 30% of the gold questions), doesn't instantly eliminate ω. Instead, based on RM's skill level estimation θ_ω, an adapted set of gold questions $\mathbb{G}^A$ are formulated by aligning their difficulty $\mathbb{G}^A_{\beta_{t_j}}$ to the corresponding worker's θ_ω, and injected within a second payload $\mathbb{P}_b$. Surpassing the correctness threshold $\mathbb{C}$ on $\mathbb{G}^A$ indicates that worker ω is indeed honest, though not up to the initial standards. As an example consider the following result from one of our experiments.

Example 1 (Correctness Threshold versus Adapted Gold Questions) assuming a correctness threshold $\mathbb{C} = 70\%$, three workers $\omega_1, \omega_2, \& \ \omega_3$ are assigned to initial payload $\mathbb{P}_a$ comprising $\mathbb{G}^B$ with difficulty levels β_{t_j} ranging between [-1.04, 1.8]. Logit values of β are interpreted accordingly: 0 is average, $\beta < 0$ implies easiness, & $\beta > 0$ implies difficulty. Given that $\omega_1^{\mathbb{C}}$ is the correctness threshold achieved by ω_1, the following correctness thresholds were achieved:
$-\omega_1^{\mathbb{C}} = 87.5\% > \mathbb{C}\,(= 70\%)$ $-\ \omega_2^{\mathbb{C}} = 50\% < \mathbb{C}\,(= 70)$ $-\ \omega_3^{\mathbb{C}} = 37.5\% < \mathbb{C}\,(= 70\%)$
Accordingly, workers $\omega_2$and ω_3 would be eliminated in a usual correctness threshold setup. In contrast, following our approach, we compute instead the workers' ability based on $\mathbb{G}^B$ upon which we formulate two $\mathbb{G}^A$, such that $\mathbb{G}_\beta^{A_{\omega_2}} \leq \theta_{\omega_2}$ and $\mathbb{G}_\beta^{A_{\omega_3}} \leq \theta_{\omega_3}$. Next ω_2, and ω_3 are assigned a second payload $\mathbb{P}_b$ comprising the respective $\mathbb{G}^A$. They scored the following correctness thresholds:
$-\ \omega_2^{\mathbb{C}} = 37.5\% < \mathbb{C}\,(= 70\%)$ $-\ \omega_3^{\mathbb{C}} = 100\% < \mathbb{C}\,(= 70\%)$
Accordingly, ω_3 is identified as a low-skilled ethical worker to be retained, unlike ω_2.

5 Gains of Recovering Low-Skilled Workers

Impact sourcing is realized through recovering low-skilled workers, who would've been otherwise treated as fraudulent and unfairly discarded. In this section, we list empirically derived heuristics for integrating low-skilled in high performing teams and illustrate how low-skilled workers' earlier payloads can be partially recovered.

5.1 High Performing Team Combinations

Following experimental results in section 6.2, three workers proved to be best as a team-size baseline. Based on a labor pool of 30 workers, 66% out of all the possible team combinations ($^{30}C_3 = 4060$ teams) produced high correctness quality (70-95%) upon aggregating their results through skill-weighted majority vote. By analyzing the teams constituting this 66%, heuristics for formulating high performing teams were empirically found. As shown below, the heuristics range from including two highly-skilled workers along with one average or low-skilled worker like $\mathbf{H_1}, \mathbf{H_2}$. Two low-skilled workers along with one highly-skilled worker like $\mathbf{H_3}, \mathbf{H_4}$. A combination of unskilled, average and highly skilled workers like $\mathbf{H_6}$**,** or average to highly skilled workers like $\mathbf{H_5}$.

Heuristics 1-6: (Heuristics for formulating High Performing Team) given a team $\mathcal{T} = \{\omega_1, \omega_2, \omega_3\}$, comprising a combination of three workers with the respective skill levels $\theta = \{\theta_{\omega_1}, \theta_{\omega_2}, \theta_{\omega_3}\}$. Logit values of θ are interpreted accordingly: 0 is average, $\theta < 0$ implies low skill level, & $\theta > 0$ implies high skill level. Through combining low-skilled with higher-skilled workers in the following team combinations, high correctness quality percentage results $\mathbb{Q}$ can be attained through skill-weighted majority vote.

– $\mathbf{H_1}$: **If** $\left(1 \le \theta_{\omega_i} < 2.5\right) \wedge (\theta_{\omega_j} < 0.6)$, **then** $65 \le \mathbb{Q} \le 95$,
  where $P(80 \le \mathbb{Q} \le 90) = 0.77$, $i = 1, 2$ and $j = 3$

– $\mathbf{H_2}$: **If** $\left(\theta_{\omega_i} \ge 0.5\right) \wedge (1 \le \theta_{\omega_j} \le 2.5)$, **then** $80 \le \mathbb{Q} \le 85$, **where** $i = 1, 2$ and $j = 3$

– $\mathbf{H_3}$: **If** $\left(-1 \le \theta_{\omega_i} < 0\right) \wedge (1 \le \theta_{\omega_j} \le 2.5)$, **then** $70 \le \mathbb{Q} \le 85$, **where** $i = 1, 2$ and $j = 3$

– $\mathbf{H_4}$: **If** $\left(-2.9 \le \theta_{\omega_i} < -1\right) \wedge (\theta_{\omega_j} > 2.5)$, **then** $55 \le \mathbb{Q} \le 80$,
  where $P(70 \le \mathbb{Q} \le 80) = 0.66$, $i = 1, 2$ and $j = 3$

– $\mathbf{H_5}$: **If**$(\theta_{\omega_i} \ge 0.5)$, **then** $70 \le \mathbb{Q} \le 80$, where $i = 1, 2, 3$

– $\mathbf{H_6}$: **If** $\left(\theta_{\omega_1} < 0\right) \wedge (\theta_{\omega_2} \ge 0.5) \wedge (\theta_{\omega_3} > 2)$, **then** $70 \le \mathbb{Q} \le 90$,
  where $P(75 \le \mathbb{Q} \le 85) = 0.78$

5.2 Recovering Partial Payloads

During the process of identifying low-skilled workers, that is, before they are assigned to form high contributing teams, low-skilled workers would've already been assigned two payloads 1) $\mathbb{P}_a$: the initial payload worker ω is assigned to, comprising balanced $\mathbb{G}^B$. Failing $\mathbb{C} > 70\%$ at this stage doesn't lead to an instant elimination, but to 2) $\mathbb{P}_b$: the second payload comprising the adapted $\mathbb{G}^A$ as per RM's computed θ_ω. This time, failing $\mathbb{C} > 70\%$ leads to elimination. Succeeding however, implies that worker ω is low-skilled and is to be henceforward enrolled to form high performing teams, which ensures high quality throughput. Rather than discarding $\mathbb{P}_a$ & $\mathbb{P}_b$, we can attain high quality results by recovering those tasks in the payloads, whose difficulty levels are within the worker's skill level.

Definition 2: (Partial Recoverable Payloads) assume a low-skilled worker ω, with computed RM's skill level θ_ω, and two payloads $\mathbb{P}_a = \{t_1^a, t_2^a, \ldots, t_m^a\}$ and $\mathbb{P}_b = \{t_1^b, t_2^b, \ldots, t_m^b\}$, where $|\mathbb{P}_a| = |\mathbb{P}_b| = m$, and have corresponding difficulty levels $\mathbb{P}_\beta^a = \{\beta_{t_1^a}, \beta_{t_2^a} \ldots, \beta_{t_m^a}\}$ and $\mathbb{P}_\beta^b = \{\beta_{t_1^b}, \beta_{t_2^b} \ldots, \beta_{t_m^b}\}$. Then the recoverable payload is:

$$\mathbb{R}_\omega^{\mathbb{P}} = \{t \in \mathbb{P}_* \mid \beta_t \le \theta_\omega\}, where\ \mathbb{P}_* = \mathbb{P}_a \cup \mathbb{P}_b, |\mathbb{R}_\omega| < 2m$$

In order to identify the recoverable tasks within a payload, their difficulty level should be computed. However, RM requires the corresponding ground truth in order to estimate the β parameter. To that end, we aim at synthesizing a reliable ground truth for the payloads' tasks, which would then serve as input to RM. We aggregate the responses of the low-skilled workers along with two other workers, such that these three workers' combination adhere to the above Heuristics 1-6 for forming a high performing team. Ultimately the skill-weighted Majority vote produce a reliable synthesized ground truth which RM uses to estimate the tasks' difficulty level. Our experiments show that the synthesized ground truth's correctness quality is always higher than 70%. We provide a description in Algorithm 1 below.

```
Algorithm 1: Recovering Partial Payloads

Input:
▪ $\mathcal{H}$ : HIT's list of questions object $q$, with attributes:1)ID: $q.ID$,
   2)difficulty: $q.difficulty$, 3)synthesized ground truth: $q.GT$
▪ $\mathfrak{T}$ : high performing team consisting of 3 workers ($\omega_1$,$\omega_2$,$\omega_3$),where $\omega_1$
   is a low-skilled worker
▪ $\mathfrak{T}^{\mathbb{C}}$ : corresponding skill levels of $\mathfrak{T}$ → ($\theta_{\omega_1}$,$\theta_{\omega_2}$,$\theta_{\omega_3}$)
▪ $\mathcal{RM}$: matrix holding list of responses of each worker in $\mathfrak{T}$
Output:
  ▪ $\mathcal{PL}$ : List of questions recovered for $\omega_1$
begin:
  for each $q$ in $\mathcal{H}$
  $q.GT$ = computeGroundTruthBySkillWeightedMajortyVote($\mathcal{RM}$, $\mathfrak{T}^{\mathbb{C}}$,$q$)
  $\mathcal{DL}$ = computeQuestionsDifficultLevelyByRaschModel($\mathcal{H}$,$\mathcal{RM}$)
  $\mathcal{ODL}$ = orderQuestionsAscendinglyByDifficultyLevel($\mathcal{DL}$)
  for each $q$ in $\mathcal{ODL}$
     if($\theta_{\omega_1}$ > $q.difficulty$)
        add $q$ to $\mathcal{PL}$
end
```

6 Experimental Results

In this section we evaluate the efficiency of adaptive gold questions in identifying low-skilled workers through laboratory and real crowdsourcing experiments. The open source eRm package for the application of IRT models in R is utilized [24], First, we investigate the percentage of low-skilled honest workers that can be correctly detected by each of $\mathbb{G}^B$ and $\mathbb{G}^A$. Next, we investigate the quality of the synthesized ground-truth from the skill-weighted Majority vote, upon which RM can estimate the task's difficulty levels, eventually allowing us to identify which parts of payloads $\mathbb{P}_a$ and $\mathbb{P}_b$ can be recovered for the correctly identified low-skilled workers. Moreover, we empirically investigate heuristics for forming high performing teams into which low-skilled workers can be later assigned to. Lastly we test our measure in a real crowdsourcing experiment.

6.1 Identifying low-skilled workers

Based on the laboratory experiment's ground truth dataset in section 3, we use the data generated from the second round, which corresponds to honest workers. This allows us to investigate how many honest workers our measure can correctly identify.

As shown in figure 3, with a correctness threshold $\mathbb{C}$=70% set, the initial payload with $\mathbb{G}^B$ retained 44.44% ethical workers (i.e. 8 out of 18). The second payload comprising $\mathbb{G}^A$ retained 50% of the previously discarded low-skilled ethical workers. That is, 72% of the honest workers have been detected after both payloads. In fact, the identified low-skilled workers get on average 90.6% of the $\mathbb{G}^A$ correctly i.e. exceeding even the original 70% correctness threshold $\mathbb{C}$ with a tangible margin. On the other hand, those ethical workers who were discarded even after $\mathbb{G}^A$ had lower skill levels than the easiest questions in $\mathbb{P}$, which justifies their exclusion.

Similarly, a laboratory based experiment comprising 30 volunteers, supports the previous findings. The initial payload with $\mathbb{G}^B$ retained 33.3% of the honest workers (i.e. 20 honest workers are discarded), while the second payload comprising $\mathbb{G}^A$ retained 65% of the previously discarded low-skilled workers (13 out of 20 discarded ethical workers were correctly retained). That is, 76% honest workers have been identified instead of 33.3%.

6.2 Investigating Crowd-synthesized Ground-truth Quality

Next, we investigate the highest crowd-synthesized ground-truth that can be attained through skill-based majority vote. A high ground-truth quality must be insured since RM base its tasks' difficulty level estimates upon it. That is, poor quality would lead to bad β estimations, which would in turn lead to wrong identification of the recoverable sections of $\mathbb{P}_a$ and $\mathbb{P}_b$. Based on the 30 honest volunteer laboratory experiment, we investigate different team combinations and search for those team combinations producing the highest ground-truth quality.

Initially, we start by all the possible combination of three-sized teams (i.e. $^{30}\mathrm{C}_3 = 4060$.) As shown in figure 4.1, many combinations: 2,671 teams i.e. ≈66%, achieve high correctness quality (70-95%). Further experiments with team combinations of 4 workers show a slight improvement, where 19,726 teams achieve correctness quality ranging between (70-95%), and 4 teams reaching 100% .i.e. ≈72% of all possible team combinations: $^{30}\mathrm{C}_4 = 27{,}405$.) On the other hand, teams of size 2 perform badly, and none reaches 95% quality.

It is clear from figure 4.1 that certain team combinations work exceedingly better

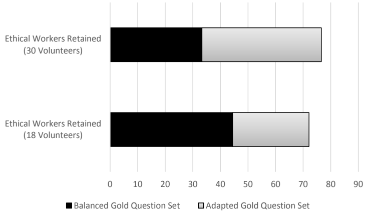
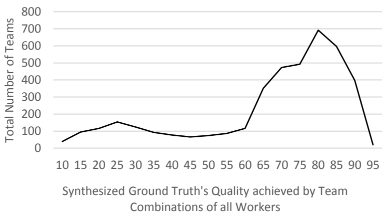
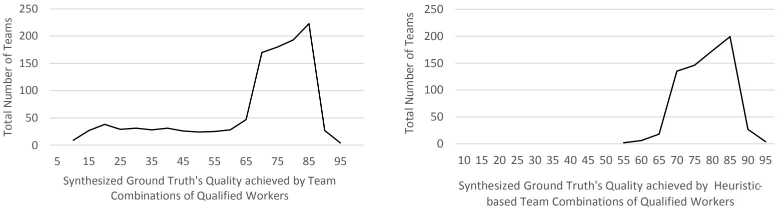


Fig. 3. Percentage of ethical workers retained after $\mathbb{G}^B$ and $\mathbb{G}^A$



4.1



4.2 **4.3**

Fig. 4. Ground truth quality achieved by different team combinations

than others. Accordingly, in figure 4.2 we zoom in only on team combinations of qualified workers (i.e. low-skilled workers that have been identified by $\mathbb{G}^A$ and highly-skilled workers who were identified earlier by $\mathbb{G}^B$). Analyzing the different skill-based team combinations producing the required high quality results (70-95%) yielded the heuristics of creating high-quality skill-based team combinations, as listed in section 5.1. Figure 4.3 depicts the Ground truth quality achieved by high performing team combinations. This yielded 718 possible team combinations, achieving on average 78% accuracy, which ranges up to 95%. Only the output of such team combinations are accordingly to be used when recovering payloads and when low-skill workers are to be integrated to form high-performing teams.

6.3 Partially Recovering Low-skilled workers' Payloads

Based on the previous experiment's findings, we check how well we can identify the recoverable sections of $\mathbb{P}_a$ and $\mathbb{P}_b$ based on RM's β estimates and the quality of the synthesized ground truth. From the 30 honest volunteer laboratory experiment, a random subset of 10 honest low-skilled workers are taken and a set of all possible high performing team combinations were created. For each worker, we compute the aggregate percentage of the recoverable payloads' size and quality over all the possible high performing team combinations this worker formulated.

As seen in figure 5, on average 68% of the payloads can be recovered (i.e. around 13 question from each of $\mathbb{P}_a$ and $\mathbb{P}_b$,). Moreover, the average correctness quality is 76%, which is even higher than the required correctness threshold. This corresponds to 6.50$ savings when recovering the initial and second payloads for each of the 10 workers, given that each payload has 20 questions and costs 50 cents.

6.4 Real crowd sourcing experiment

We evaluate the efficiency of our measure through a real crowdsourcing experiment, which was ran on the generic CrowdFlower crowd sourcing platform. A total of 41 workers were assigned Hits comprising 10 payload questions and 4 gold questions, giving a total of 574 judgments and costing 20.5$, where each HIT costs 35 cents. A correctness threshold $\mathbb{C}$ = 70% would discard around 30% of the workers (i.e. 12 workers). In contrast, our measure, assigned those 12 workers to a second payload $\mathbb{P}_b$

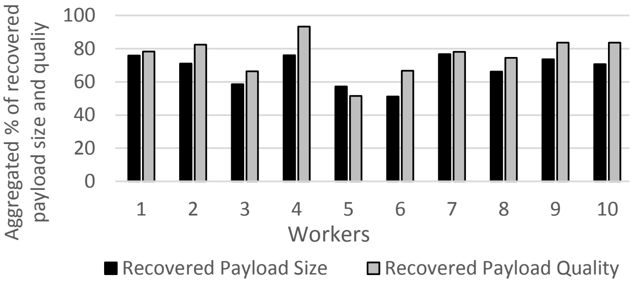


Fig. 5. Size and Quality of Recovered Payloads

comprising adapted $\mathbb{G}^A$. This yielded a total of 168 judgments, costing 4.2$. In total, 25% of the workers were identified as low-skilled workers (i.e. 3 workers.)

Unlike the laboratory-based experiment, the real crowd sourcing experiment has no ground truth (i.e. number of low-skilled workers and number of fraudulent workers), accordingly we measure the efficiency of how well these workers were correctly identified by checking the quality of their partial recovered payloads, since these payload tasks are those within their real skill level. On average 50% of both payloads $\mathbb{P}_a$ and $\mathbb{P}_b$were recovered with an average correctness of 80%, which surpasses even the correctness threshold. This corresponds to 3 payloads (i.e. 1.5$). The small savings reflect nothing more than the number of detected low-skilled workers, whose percentage in this experiment could have been small and lesser than the fraudulent workers.

7 Summary and Future Work

In this paper, we support *Impact Sourcing* by developing a socially responsible measure: adaptive gold questions. Our laboratory-based experiment attests that current employed quality control measures like gold questions or reputation based systems tend to misjudge low-skilled workers and eventually discard them from the labor pool. In contrast, we show how gold questions that are adapted to the corresponding workers' ability can identify low-skilled workers, consequently saving them from the vicious elimination cycle and allow them to work within their skill levels. This can be achieved by utilizing the Rasch Model, which estimates and aligns both the workers' skill level and the gold questions' difficulty level. Furthermore, we show how initial payloads could be partially recovered to reclaim some of the arguable economic loses. Through empirical results, we defined heuristics for building high performing teams. Following these heuristics, low-skilled workers can be effectively integrated to produce reliable results (70-95%) through skill-weighted majority vote.

Nevertheless, retaining a database of workers and dynamically creating such high performing teams might not always be feasible. Therefore, the next step would be to expand our model's adaptivity to encompass not only the gold questions, but to adapt as well the entire payload to suit each workers' ability, which would boost the overall quality and promote a more efficient assignment of tasks.

8 References

[1] "Digital Jobs Africa: The Rockefeller Foundation," [Online]. Available: http://www.rockefellerfoundation.org/our-work/current-work/digital-jobs-africa/impact-sourcing.

[2] J. Selke, C. Lofi, and W.-T. Balke, "Pushing the Boundaries of Crowd-Enabled Databases with Query-Driven Schema Expansion, " *in 38th Int. Conf. VLDB,* 2012, pp. 538-549.

[3] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating Named Entities in Twitter Data with Crowdsourcing," *CSLDAMT '10 Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 80–88, 2010.

[4] C. Lofi, K. El Maarry, and W.-T. Balke, “Skyline Queries in Crowd-Enabled Databases,” *EDBT/ICDT Joint Conf., Proc. of the 16th Int. Conf. on Extending Database Technology* 2013.

[5] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” *International AAAI Conf. on Weblogs& Social Media*, pp. 538–541, 2011.

[6] C. Callison-Burch, “Fast, cheap, and creative: evaluating translation quality using Amazon’s Mechanical Turk,” *EMNLP’09: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, vol. 1, no. 1, pp. 286–295, 2009.

[7] C. Lofi, J. Selke, and W.-T. Balke, “Information Extraction Meets Crowdsourcing: A Promising Couple,” *Proc. of the VLDB Endowment 5 (6), 538-549, 2012. 23, 2012*.

[8] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, “Limits on the majority vote accuracy in classifier fusion,” *Journal: Pattern Analysis and Applications - PAA ,* vol. 6, no. 1, pp. 22-31, 2003

[9] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the EM algorithm,” *Journal of Applied Statistics.* vol. 28, pp. 20–28, 1979.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, “Learning From Crowds,” *The Journal of Machine Learning Research* , vol. 11, pp. 1297–1322, 2010.

[11] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, “Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise,” *In Proc. of NIPS,* vol. 22, no. 1, pp. 1–9, 2009.

[12] P. G. Ipeirotis, F. Provost, and J. Wang, “Quality Management on Amazon Mechanical Turk,” *Proc. of ACM SIGKDD Workshop on Human Computation*, 2010, pp. 0–3.

[13] K. El Maarry, W.-T. Balke, H. Cho, S. Hwang, and Y. Baba, “Skill ontology-based model for Quality Assurance in Crowdsourcing,” *UnCrowd 2014: DASFAA Workshop on Uncertain and Crowdsourced Data, Bali, Indonesia*, 2014.

[14] A. Ignjatovic, N. Foo, and C. T. L. C. T. Lee, “An Analytic Approach to Reputation Ranking of Participants in Online Transactions,” *2008 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 1, 2008.

[15] Z. Noorian and M. Ulieru, “The State of the Art in Trust and Reputation Systems: A Framework for Comparison,” *Journal of theoretical and applied electronic commerce research,* vol. 5, no. 2. 2010.

[16] G. Kazai, “In Search of Quality in Crowdsourcing for Search Engine Evaluation,” *ECIR'11: Proc. of the 33rd European conf. on Advances in information retrieval*, vol. 44, no. 2, pp. 165–176, 2011.

[17] R. E. Traub, “Applications of item response theory to practical testing problems,” *Book's Publisher: Erlbaum Associates,* vol. 5, pp. 539–543, 1980.

[18] G. Rasch, “Probabilistic Models for Some Intelligence and Attainment Tests,” *Book's Publisher: Nielsen & Lydiche*, 1960.

[19] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer, “An Evaluation of Aggregation Techniques in Crowdsourcing,” *WISE* 2013.

[20] J. Wang, P. G. Ipeirotis, and F. Provost, “Managing Crowdsourced Workers,” *Winter Conf. on Business Intelligence*, 2011.

[21] W. H. Batchelder and A. K. Romney, “Test theory without an answer key,” *Journal Psychometrika*, Volume 53, Issue 1, pp. 71–92, 1988.

[22] K. El Maarry, and W.-T. Balke, “Retaining Rough Diamonds: Towards a Fairer Elimination of Low-skilled Workers,” *20th Int. Conf. on Database Systems for Advanced Applications (DASFAA), Hanoi, Vietnam,* 2015.

[23] G. Karabatsos, “A critique of Rasch residual fit statistics,” *Journal of Applied Measures.*, vol. 1, no. 2, pp. 152–176, 2000.

[24] P. Mair, “Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R,” *Journal of Statistical Software,* vol. 20, no. 9, pp. 1-20, 2007.