Retaining Rough Diamonds: Towards a Fairer Elimination of Low-skilled Workers

Kinda El Maarry, Wolf-Tilo Balke

Institut für Informationssysteme, TU Braunschweig, Germany
{elmaarry, balke}@ifis.cs.tu-bs.de

Abstract.
Living the economic dream of globalization in the form of a location- and time-independent world-wide employment market, today crowd sourcing companies offer affordable digital solutions to business problems. At the same time, highly accessible economic opportunities are offered to workers, who often live in low or middle income countries. Thus, crowd sourcing can be understood as a flexible social solution that indiscriminately reaches out to poor, yet diligent workers: a win-win situation for employers and crowd workers. On the other hand, its virtual nature opens doors to unethical exploitation by fraudulent workers, compromising in turn the overall quality of the gained results and increasing the costs of continuous result quality assurance, e.g. by gold questions or majority votes. The central question discussed in this paper is how to distinguish between basically honest workers, who might just be lacking educational skills, and plainly unethical workers. We show how current quality control measures misjudge and subsequently discriminate against honest workers with lower skill levels. In contrast, our techniques use statistical models that computes the level of a worker's skill and a task's difficulty to clearly distinguish each worker's success zone and detect irrational response patterns, which usually imply fraud. Our evaluation shows that about 50% of misjudged workers can be successfully detected as honest, can be retained, and subsequently redirected to easier tasks.

Keywords: crowd sourcing, impact sourcing, fraud detection

1 Introduction

"It's been a dream, me having my own place, paying my own rent, buying my own food. Being independent," says Martha, a Samasource Kenyan worker[1]; one of the many faces behind the international taskforce ready to work through crowd sourcing platforms. The social model of *Impact Sourcing*, first implemented by the social enterprise Digital Divide Data (DDD)[2] back in 2001 has been adopted by many compa-

[1] http://www.samasource.org/impact/.
[2] http://www.digitaldividedata.com/about/

nies and crowd sourcing platforms like Samasource[3], RuralShores[4], etc. The new Impact Sourcing industry aims at hiring people at the bottom of the income pyramid to perform small, yet useful cognitive and intelligent tasks via digital interfaces, which ultimately promises to boost the general economic development [1]. However, the mostly anonymous, highly distributed and virtual nature of the short-term work contracts also carry the danger of being exploited by fraudulent workers: By providing incorrect (usually simply random) answers, they compromise the overall result quality and thus not only directly hurt the respective task provider, but in the long run also all honest workers in dire need of employment.

Anecdotic evidence can be drawn from our own research work on crowd sourcing as reported in [2]: By completely excluding workers from just two offending countries, where the number of clearly fraudulent workers seemed to be much higher than average, the overall result correctness in our experiments instantly increased by about 20%. In particular, correctness for a simple genre classification task for movies increased from 59% to 79% using majority vote for quality assurance. Of course, this crude heuristics was bound to exclude many honest workers too.

Although typical quality control measures for crowd sourcing like gold questions or majority votes promise to mitigate this problem, they also face serious problems: firstly, they are only applicable for factual tasks, which hampers creative task design and thus overall benefit. Secondly, they incur additional costs for the task provider and thus reduce the readiness to crowd source tasks, and finally (and probably worst for impact sourcing) they exclude honest and willing workers that may not have been provided tasks on their individual skill levels. Indeed, the bottom of the income pyramid encompasses a heterogeneous set of workers, who according to their skill level provide responses that are: either *sufficiently good* by non-expert, diligent workers with higher skill levels, or *mixed* by honest workers with lower skill levels, as well as by unethical workers who exploit the system for financial gains. Yet, according to Samasource 92% of its taskforce are unemployed or underemployed, i.e. for the crowd, every incoming living wage counts and contributes to a better standard of living.

On the other end of the spectrum, crowd sourcing emerges as an unparalleled solution [3] to companies with intelligent digital tasks, to name but a few: text translation, image tagging, text sentiment analysis. Generally speaking, through the collective intelligence of the diligent workers, high quality responses could be attained. Naively, such high quality can be assured by manually cutting out the unethical workers through submitted response checks. However, this instantly invalidates the core gains attained through crowd sourcing, and becomes both costly and time consuming. Consequently, unethical workers are further encouraged to submit low quality results, and it becomes a question of automatically detecting such workers for an improved overall quality [4].

As argued above, some common practices are 1) injecting a set of questions whose answers are already known, so-called gold questions, within each Human Intelligent Task (HIT), 2) Employing Majority vote to filter out workers who often fail to agree

[3] http://www.samasource.org/.
[4] http://ruralshores.com/about.html.

with the majority, or 3) adopting a reputation based system, where workers' history is recorded, and a reputation score is assigned to each worker by combining e.g., the requestor's satisfaction level, the ratio of completed to aborted HITS, etc. Unfortunately, such practices can heavily misjudge honest, yet less skilled workers.

Example 1 *(Collecting motion picture ratings):*
Given a dataset of movies, the crowd is asked to classify each movie as either PG, or PG-13. Assume a skewed distribution where 90% of the movies are PG, and only 10% are PG-13. Assume worker A simply tags all movies as PG, ultimately he/she will only have a 10% error rate. On the other hand, consider worker B who's actually checking each movie. Worker B can easily exhibit similar or even higher error rates, because he/she perceives some movies according to his/her standards as PG-13. Although worker A is obviously a spammer, in a reputation based system he/she would be given a higher reputation score than worker B, since more gold questions were answered correctly.

Our results in initial experiments indeed indicate that even with datasets where answer possibilities are evenly distributed, unethical workers can still through random guessing surpass honest workers with lower skill levels. Based on these insights, in this paper we design a method that abstracts each worker's gold questions' responses to a skill-graded response vector, and then zooms in on their individual success zone for quality control. The success zones are individually defined by each worker's skill level and bounded by questions' whose difficulty levels are well within the worker's skill level. We can compute both parameters through psychometric *item response theory* (IRT) models: in particular, the Rasch model [5]. The underlying assumption is that honest workers should exhibit a clear tendency to correctly answer tasks within their difficulty levels. Moreover, failing to answer easy tasks, yet correctly answering more difficult ones would indicate fraud. We develop three techniques that are designed to detect irrational response patterns in success zones. The contributions of this paper can be summarized as follows:

- We show how gold questions and reputation-based systems can be *bypassed by unethical workers*, using real-world dataset in a laboratory-based study.
- We present a framework for distinguishing between unethical and honest workers with lower skill levels in *a fair, yet reliable* manner.
- We extensively test our framework in practical crowd sourcing experiments and demonstrate how honest workers with lower skills levels can be indeed detected and redirected to skill-fitted tasks.

The rest of the paper is structured as follows: In section 2, we give an overview of related work. In section 3, we motivate our problem with a case study, then start describing our framework in section 4, by presenting the underlying statistical Rasch model and illustrating how it can be used to identify workers' success zone. In section 5, we introduce three techniques that aim at recognizing irrational patterns in success zones. This is backed up by a laboratory-based experiment that offers ground-truth and a real-world crowd sourcing experiment in the evaluation section. Finally, the last section gives a summary and an overview of future work.

2 Related Work

Crowdsourcing provides both a *social* chance that indiscriminately reaches out to poor, yet diligent workers, as well as an affordable digital solution for companies with intelligent digital business problems like e.g., web resource tagging [6], completing missing data [7], sentiment analysis [8], text translation [9], information extraction [10], etc. But as with every chance, the challenge of acquiring high quality results, which is compromised by unethical workers, must be overcome. In this section, we give an overview of the current crowd sourcing quality control measures, as well as a brief overview of the Rasch Model and its related work in crowd sourcing.

A rich body of research has examined many different techniques to mitigate the quality problem in crowdsourcing. Aggregation methods aim at improving the overall quality through redundancy and repeated labeling. Through assigning several workers to the same task, an aggregation of their responses help identify the correct response. Such aggregation methods include: basic Majority decision, which has been shown to have severe limitations [11]. This was further developed by Dawid and Skene [12] to take the response's quality based on the workers into consideration, through applying an expectation maximization algorithm. Other approaches that considered such error rates relied on: Bayesian version of the expectation maximization algorithm approach [13], or a probabilistic approach that takes into account both the worker's skill and the difficulty of the task at hand [14]. A further step was taken in [15] with an algorithm separating the unrecoverable error rates from recoverable bias. Manipulating monetary incentives have also been investigated, yet proves tricky to implement, where low paid jobs yield sloppy work, and high paid jobs attract unethical workers [16].

Other techniques focus on trying to eliminate unethical workers on longer time scales like constantly measuring performance with injected gold questions or employing the workforce via reputation-based systems. Even when the procurement of gold questions is feasible and not too expensive, the question of how many gold questions to include immediately materializes [17]. On the other hand reliably computing the workers' reputation poses a real challenge, and many reputation approaches have been investigated whether it's based on a reputation model [18-19], on feedback and overall satisfaction [19], or on deterministic approaches [20], etc.

Our work is tightly related to the IRT paradigm [21] in psychometrics, which enables us to focus on the workers' capabilities. More specifically, we employ the Rasch models [5], which computes the expected response correctness probability of a worker to a given task, the task's difficulty, and the ability of the worker. This helps us address a principal concern of Impact sourcing: distinguishing honest workers with low skill levels from unethical workers. So far, most research have focused on one or two of those aspects, with the exception to the work of Whitehill in [14], where they presented GLAD – a generative model of labels, abilities and difficulties. The presented model is close to our work as it's also based on IRT. They obtain the maximum likelihood estimates of these three parameters through utilizing an Expectation-Maximization approach (EM) in an iterative manner for results aggregation. GLAD's robustness wavers when faced with unethical workers, especially when they constitute more than 30% of the task force [22]. Our perspective is however focused on detecting irrational response patterns to be able to distinguish diligent workers with lower skill levels.

Other works, considered only one aspect, namely the worker's ability. Dawid and Skene [12] utilized confusion matrices, which offered an improved form of the redundancy technique. However, as pointed out and addressed by Ipeirotis [23], this underestimates the quality of workers who consistently give incorrect results. In contrast, in [24], the workers' ability together with the inference of correct answers are investigated. The downside however, is that it overlooks the varying difficulties of the task at hand, which should influence the workers' abilities. In our model, the correctness probability of a worker for a given task isn't static, but varies according to the task's difficulty and the worker's ability. Furthermore, it's measured across the different tasks' difficulty level.

3 Motivational Crowd sourcing Laboratory-based Study

To acquire a basic ground truth dataset, we conducted a small-scale laboratory-based experiment, where a total of 18 workers volunteered for the study. Given a set of 20 multiple choice questions, the volunteers were asked to answer the questions twice. In the first round, they would randomly select answers in any fashion. In the second round, they were asked to truthfully consider the questions before answering.

In this paper we formulate our HITs over an American standardized test for college admission with medium difficulty level: the Graduate Record Examination (GRE) dataset, which was crawled from graduateshotline.com. A GRE test is typically made up of five sections. We extracted the questions corresponding to the verbal section, namely the verbal practice questions. The task then is to select the right definition of a given word. For each question, four definitions are given, and the worker should select the correct definition corresponding to the word in question.

Figure 1 sorts all workers' answer sets according to the respective total number of correct answers achieved over the 20 questions. Although no worker got all 20 answers right, it comes as no surprise that truthful answers tend to be more correct than

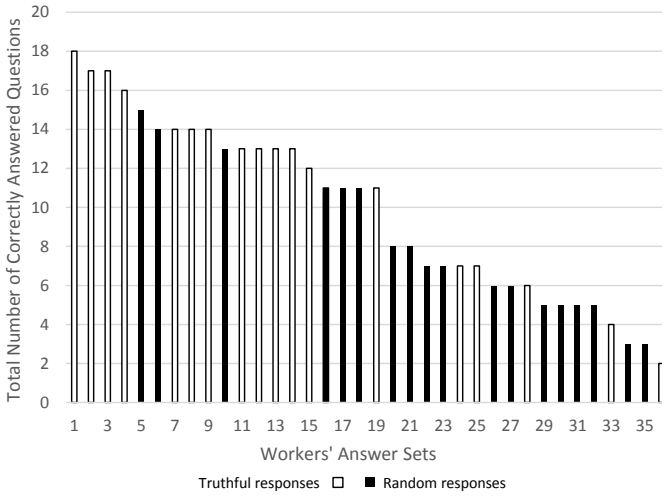

Fig. 1. Truthful versus Random Responses

random answers: in the random response round, the workers had on average 40% correct answers, while in the truthful response round, the workers had on average 58.6% correct answers. Furthermore, one can clearly see that even though the dataset is in no way biased, random responses at times produced better overall results than workers who actually tried to truthfully answer the questions.

3.1 Unethical workers versus Reputation-based systems.

Consider the top ten 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, who scored only 14 correct answers. Yet, here three workers at least tried to answer correctly. In biased datasets, (e.g. the movie dataset from example 1), experienced unethical workers' could easily improve their reputation score even more by choosing proper answering schemes.

3.2 Unethical workers versus Gold Questions.

Upon setting up 40% gold questions (i.e. 8 gold questions out of the 20 questions) and a 70% accuracy level threshold (i.e. workers scoring less than 70% correct answers in the gold questions are eliminated), we considered the following three gold question set scenarios as depicted in Figure 2:

1. *Easy gold question set:* 38.8% honest workers discarded (i.e. 7 workers) and 66.67% unethical workers eliminated (i.e. 12 workers).
2. *Balanced gold question set, addressing all difficulty levels*: 61% honest workers discarded (i.e. 11 workers are eliminated) and 88% random workers eliminated (i.e. 16 workers are eliminated).
3. *Difficult gold question set*: 55.5% honest workers discarded[5] (i.e. 10 workers) and 88.8% unethical workers eliminated (i.e. 16 workers).

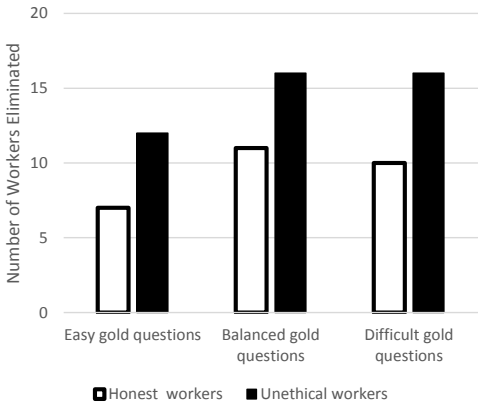

Fig. 2. Number of workers Eliminated given different types of gold questions

Although gold questions are more biased to penalize unethical workers, still this bias is relatively low, and a significant number of honest workers are penalized, too. Moreover, unethical workers can with a higher chance bypass an easy set of gold questions (33.33% unethical workers kept) than they are to bypass either a balanced or a difficult set of gold questions (around 12% unethical workers kept).

[5]Typically, it's to be expected that the number of honest workers eliminated by the difficult set of gold questions would be higher than that with the balanced gold, as it would impose a higher skill threshold on the employed workers. However, due to the relatively small number of volunteers, the 6% difference which is in fact only 1 worker more that was eliminated by the balanced gold question can be misleading and should be considered as an artifact of the experiment and not generalized. For the rest of our experiments, we use a balanced set of gold questions.

4 Identifying Success Zone Bounds

In this section we provide a short overview of the underlying statistical *Rasch Model* (RM), and illustrate how we employ it to 1) abstract workers' responses to a skill-graded vector, and 2) identify each worker's success zone for quality control.

4.1 The Rasch Model

Crowdsourcing inherently involves many human factors, accordingly our attention is drawn to the science assessing individual's capabilities, aptitudes and intelligence, namely, psychometrics and its IRT classes, in particular the RM. Simply put, the RM computes the probability P_{ij} that a worker's $w_i \in W$ response to a given task $t_j \in T$ is correct as a function of his ability θ_{w_i} and the difficulty of the task β_{t_j}. Assuming a binary setting, where a worker's response $x_{ij} \in [0,1]$ is known (with 0 meaning an incorrect and 1 a correct response), RM's dichotomous case can be employed. Basically, 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 $W= \{w_1, w_2, \dots, w_n\}$, where $|W| = n$, and a HIT $T= \{t_1, t_2, \dots, t_m\}$, where $|T| = m$. Assume $w_i \in W$ and $t_j \in T$, then correctness Probability P_{ij} can be given as follows

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

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

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

Such a logit scale enforces a consistent valued unit interval that is meaningful between the locations of the workers and the questions when they're drawn on a map [25]. So whereas a worker w_i has a 50% chance of answering a question within

his/her exact ability, this success probability increases to 75% for questions that is 1 logit easier and similarly drops to 25% for questions that is 1 logit more difficult. The difficulty of a question with a logit value of 0 is average. A negative logit value implies an easy β_{t_j} and a low $\theta_{w_i}$and vice versa. Accordingly, the correctness probability of a worker's response will be high when his/her ability exceeds the corresponding task's difficulty. We exploit this to detect irrational response patterns in section 5.3.

Objectively Estimating θ and β Parameters.
Naively, β can be determined by observing the proportion of incorrect responses, while θ_{w_i}could be defined by where a worker w_i stands in the percentile among all the other workers. In fact, the practice of using the raw score (i.e. total number of correct responses) to estimate the worker's ability is quite ubiquitous [25]. However, this fails to capture the reliability of responses and raises many questions as can be seen in the following example:

Example 2:
a) Assume that the same HIT is assigned to two different groups of workers. The majority of the first group's workers are unethical, while the majority of the second is honest. A non-objective measure of β would label questions assigned to the first group as difficult, because the majority gave incorrect responses. Simultaneously, the same questions would have a lower β for the second honest group of workers.

b) Assume two HITs with different β, where the first has a low difficulty level and the second has a high difficulty level. An unobjective measure of θ would equally treat the workers lying in the same percentile of each HIT, irrespective of the difficulty of the questions they're handling.

Accordingly, objectivity is needed. Simply put 1) θ measurements should be independent of *T*, and 2) β measurements should be independent *W*. In RM, the "objective" measurement of (θ, β) is emphasized [26].

Numerous ways for estimating the Rasch model's parameters (θ, β) exist. We employ the conditional maximum likelihood estimation (CML) for estimating β, as it leads to minimal estimation bias with well-defined standard error. For estimating the θ, join maximum likelihood estimation (JML) is used, which proved to be robust against missing data, as well as an efficient estimation method [27].

These parameters can only be estimated on workers' responses to gold questions, where all responses can be judged to be correct or not. In the evaluation section, we illustrate that 40% gold question suffice for the RM to correctly approximate each of the (θ, β) parameters (i.e. for a HIT of 20 questions, 8 gold questions are needed).

4.2 Skill-graded Vectors and Success Zones

In a perfect setting, a worker who's able to correctly answer questions of a certain difficulty level, should be able to answer all less difficult questions. Similarly, failing to submit a correct response to a question of a certain difficulty would imply the worker's incompetency in answering more difficult questions. Different workers will

have different skill levels and a fair judgment of a worker should be based on how well they do on tasks within their skill level. To that end, we reorder each worker's gold question response vectors, such that the questions are ordered ascendingly in terms of their difficulty using the RM's estimated β parameter. At which point, in an ideal setting, we can observe the following workers' response vectors as illustrated in Table 3. Next, we zoom in on the set of questions within the worker's ability (i.e., success zone). Using the RM's estimated θ parameter, we can define the bound of the success zone.

Table 1: (Ideal) Skill-graded Response Matrix

Tasks / Workers	1	2	3	4	5	6	7	8	9
A	1	1	1	0	0	0	0	0	0
B	1	1	1	1	1	0	0	0	0
C	1	1	1	1	1	1	1	0	0

Definition 2: *(Skill-graded Vector -* ***SV)*** for w_i assigned to a HIT with golden questions $T = \{t_1, t_2, \dots, t_m\}$, $|T| = m$, with corresponding $\beta = \{\beta_{t_1}, \beta_{t_2}, \dots, \beta_{t_m}\}$. Worker w_i submits the following corresponding response vector $RV_{w_i} = \{x_1, x_2, \dots, x_m\}$.

$$SV_{w_i} = \{\{x_a, x_b, \dots, x_z\} \mid \beta_{t_z} > \cdots > \beta_{t_b} > \beta_{t_a}\},\ where\ |SV_{w_i}| = |T| = m$$

Definition 3: *(Success Zone –* ***SZ)*** *given* a Skill-graded Vector $SV_{w_i} = \{x_1, x_2, \dots, x_m\}$ for worker w_i to a set of gold questions $|T| = m$.

$$SZ_{w_i} = \{\{x_a, x_b, \dots, x_h\} \mid h \leq m \wedge \beta_{t_{h+1}} > \theta_{w_i}\}$$

5 Rational and Irrational patterns in Success Zones

In reality, the SV matrix shown in Table 1, also known as a perfect Guttmann scale [28], is more plausible in theory, and is rather the exception than the rule. Observing the workers in reality, different response patterns can be seen, which are trickier to understand and handle. Table 2 illustrates a more realistic SV matrix, the shaded cells represent the success zones of each worker. Workers will sometimes miss out on easy questions within their ability ($\theta_{w_i} > \beta_{t_j}$) e.g. worker A responds incorrectly to task 3. And sometimes they may as well successfully respond to difficult questions beyond their ability ($\theta_{w_i} < \beta_{t_j}$) e.g. worker C answers task 8 correctly. Some of these different response patterns can be explained [26] as seen in Table 3.

Table 2: (Realistic) Skill-graded Response Matrix

Workers \ Tasks	1	2	3	4	5	6	7	8	9
A	1	1	0	1	1	1	1	0	0
B	1	1	1	0	1	0	0	1	0
C	1	0	1	1	1	0	0	1	0

Accordingly, there will be unexpected false or correct responses in a worker's SV. These discrepancies are however of greater or lesser importance and impact the overall rationality of the response pattern depending on the number of their occurrences and their location in SV. We only focus on discrepancies within success zones, while SV entries that are outside of the success zone are: 1) incorrect and workers shouldn't be penalized for or, 2) correct and will be often attained through guessing.

Next we present three techniques that focus on success zones and aim at recognizing irrational patterns within these zones: 1) Skill-adapted Gold questions, 2) Entropy-based elimination, and 3) Rasch-based elimination.

Table 3: Observed response patterns

Response Pattern	Response Vector
Lucky-guessing	A few unexpected correct responses to difficult tasks
Carelessness	A few unexpected incorrect responses to easy tasks
Rushing	Unexpected error near the end for tasks with difficulty levels within a worker's abilities
Random guessing	Unexpected correct and incorrect responses irrespective of the question's difficulty
Constant response set	Same submitted response over and over again

5.1 Skill-adapted gold questions

This technique is similar to using gold questions with a certain accuracy level (i.e. workers not achieving the required accuracy level are eliminated), except that gold questions don't provide a fair basis for judging workers with varying skill levels. The new technique also applies the gold question's accuracy level, but strictly on the success zone i.e. it discards all workers failing to achieve the required accuracy level on tasks within their skill level. Accordingly an irrational pattern in a worker's success zone can be defined as follows:

Definition 4: *(Skill-adapted gold questions)* given a success zone $SZ_{w_i} = \{x_1, x_2, \dots, x_h\}$ of a worker w_i and a required accuracy level A, all workers are eliminated for which holds:

$$\sum_{j=0}^{h} x_j \Big/ |SZ_{w_i}| < A$$

5.2 Entropy-based Elimination

We utilize the *Shannon Entropy* [29] to measure the randomness of responses within the success zone. Following Shannon's entropy definition, a response pattern with high entropy indicates randomness, while rational response pattern will have low entropy. According to our experimental results on our dataset, entropy values higher than 0.89 indicated randomness. Accordingly an irrational pattern in a worker's success zone can be defined as follows:

Definition 5: *(Entropy-based Elimination)* given a success zone $SZ_{w_i} = \{x_1, x_1, \dots, x_h\}$ of a worker w_i and an entropy threshold=0.89, all workers are eliminated for which holds:

$$-\sum_{j=0}^{h} p(x_j)\ \log_2 p(x_j) > 0.89$$

5.3 Rasch-based Elimination

One shortcoming of both the skill-adapted gold questions and the entropy-based elimination technique, is that none of them take into consideration the point at which an incorrect response is given. For instance, the following two response vectors (101011) and (111100) get the same entropy of 0.9182. Following RM's definition of having higher correctness probability the easier the task is relative to a worker's skill ability, then failing on the second task in the first response pattern should be penalized more than failing on the fifth task in the second response pattern. Using the correctness probability P_{ij} that is estimated by the Rasch model, we add up the correctness probabilities for task entries that are correctly answered and penalize falsely answered tasks by subtracting its corresponding correctness probability. We then define an irrational response pattern as follows:

Definition 6: *(Rasch-based Elimination)* given a success zone $SZ_{w_i} = \{x_1, x_1, \dots, x_h\}$ and the corresponding computed correctness probabilities $P = \{P_{i1}, P_{i2}, \dots, P_{ih}\}$ of a worker w_i and a required accuracy level A, all workers are eliminated for which holds:

$$\sum_{j=0}^{h} \left(-P_{ij} + (x_j * 2P_{ij})\right) < A$$

6 Experimental Results

In this section we evaluate the proposed framework and extensively evaluate the efficiency of the three proposed techniques: 1) Skill-adapted Gold Questions, 2) Entropy-based elimination and 3) Rasch-based elimination.

In section 3, we evaluated the impact of the type of gold questions used (easy, balanced and difficult). Next, we investigate how many gold questions should be injected, which would allow the rasch model to correctly approximate each worker's skill level, and their correctness probability. Similar to the motivational case study in section 3, we use the verbal section from the GRE dataset to create our HITS for evaluation purposes.

The open source eRm package for the application of IRT models in R is utilized, where correct responses are coded as 1 and incorrect ones are coded as 0 in a worker's success zone. The eRm package uses conditional maximum likelihood CML estimation as it maintains the concept of specific objectivity [5], [30].

6.1 Gold question set size

We experiment with different gold question set sizes (κ) and examine how it impacts the reliability of the RM's approximated parameters (θ_{w_i} worker's skill level and P_{ij} worker's correctness probability). This ultimately allows the inference of a heuristic for gold question set size κ, which would permit a good a reliable approximation of the parameters that are used in identifying a worker's success zone. Starting off with $\kappa = 20\%$ (i.e. 4 gold questions), we incrementally upsurge the size till 80%. We designed the different sets to be balanced in terms of questions' difficulty. Figure 3

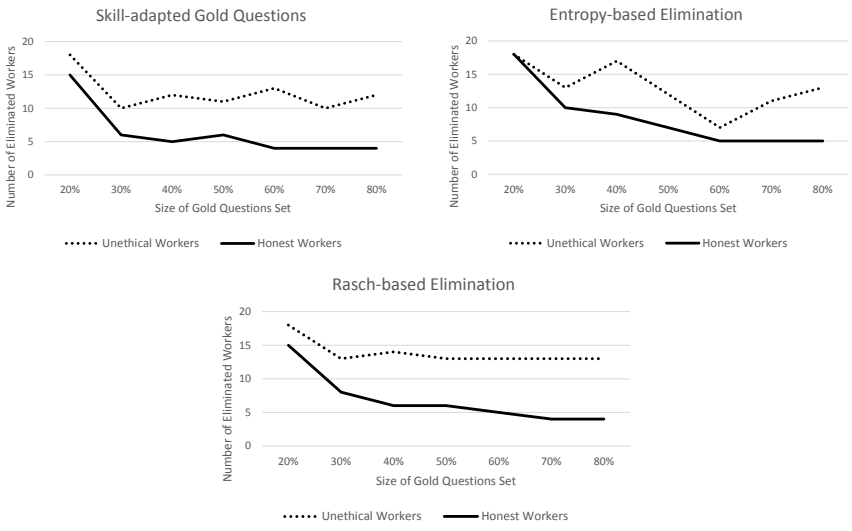

Fig. 3. Impact of varying κ

illustrates the impact of the different κ sizes on the three techniques in terms of: 1) number of eliminated unethical workers and 2) number of misjudged honest workers.

As κ increases, the three techniques can more efficiently judge honest workers and the number of misjudgments decrease. Moreover, for the three techniques the number of misjudged honest workers curves tend to converge on average when κ = 40%.

Both the skill-adapted gold questions and the Rasch-based elimination approaches are optimistic compared to the Entropy-based elimination, which maintains on average a higher rate of eliminated unethical workers. Interesting to note is that convergence is much smoother and pronounced with the number of misjudged ethical workers curve than that of the unethical workers. This can be attributed to the implicit random response nature of the unethical workers i.e. for every set of gold questions that are of the same size, the unethical workers' performance would vastly fluctuate. This is especially prominent with the pessimistic entropy-based elimination approach.

For the rest of our experiments, we design our HITS to constitute 40% gold questions.

6.2 Ground truth-based evaluation

We use the dataset from the laboratory-based study in section 3, which is made up of 20 questions, 40% out of which are gold questions (i.e. 8 questions). This gives us ground truth, where responses corresponding to the first random round constitutes unethical workers, while the second truthful response round constitutes honest workers. Accordingly, we can precisely evaluate how each of the proposed techniques fare in terms of detecting and distinguishing between irrational and rational response patterns in success zones and accordingly eliminate the workers fairly.

Figure 4 depicts the number of eliminated workers by each of the: a) skill-adapted gold questions, b)entropy-based elimination, c)Rasch-based elimination techniques, and compares them to the standard gold questions technique. For the gold question technique, we use a balanced set (i.e. set of gold questions target all difficulty levels) with an accuracy level set to 70% (i.e. workers must provide 70% correct answers, otherwise they're eliminated).

As can be seen, the three proposed techniques perform significantly better than solely relying on gold questions. Both the Skill-adapted gold questions and the Rasch-based elimination retain 50% more of the honest workers. Whereas gold questions misjudge 61% of the honest workers, both the skill-adapted gold questions and the Rasch-based elimination misjudge only 22% and 33% honest workers, respectively. Moreover, in support of our earlier findings, the elimination ratio of unethical to honest workers for the skill-adapted gold questions and the Rasch-based elimination is around 40%, designating them as more optimistic techniques. On the other hand, the entropy-based elimination is more rigid and pessimistic with a 52% ratio, yet still much better than simple gold questions. So while both the Skill-adapted gold questions and Rasch-based elimination can more efficiently recognize honest workers, entropy-based elimination tends to more efficiently recognize unethical workers. In summary, all three techniques exhibit better ratios than gold questions.

A closer look on the eliminated honest workers by skill-adapted gold questions techniques ascertains the justification of these elimination, where eliminated workers

either: a) failed in answering all gold questions or b) failed in having a skill level higher than the easiest question (e.g. worker abilityθ_{w_i} = -1.24 and the easiest question's difficulty level β_{t_1}= -1.21).

Furthermore, examining the different computed entropies for both honest and unethical workers, we observed that entropies higher than 0.89 indicated randomness in the responses. Accordingly, by setting the entropy threshold to 0.89, 94% unethical workers are eliminated, and 50% ethical honest are misjudged.

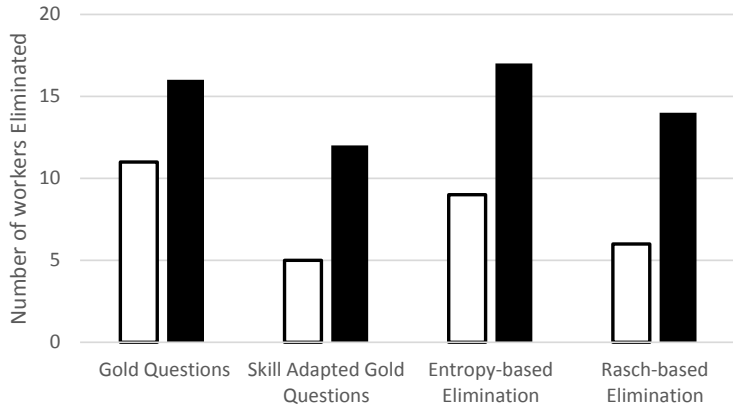

Fig. 4. Number of Ethical/Unethical Workers Eliminated

6.3 Practical crowd sourcing experiments on real world data

In this experiment, we evaluate the efficiency of our techniques in real world crowd sourcing experiments. We used CrowdFlower as a generic crowd sourcing platform. Following the results of sections 3 and 6.1, we designed a balanced set of Gold questions. Each HIT consisted of 28 questions. We overrode the gold question policy of CrowdFlower, allowing workers to participate in the job, even if they did not meet the required accuracy level. No restraints were set to geography, and skills were set to minimum as defined by the platform. We collected 1148 judgments from 41 workers incurring a total cost of 35$. Table 4 illustrates how many workers were eliminated by each technique.

The results closely reflect our results on synthetic data. Using the platform's static accuracy levels, 32% of the workers would have been eliminated (13 workers). Both the Skill-adapted gold questions and Rasch-based Elimination tend to retain 50% of the workers originally eliminated by gold questions. On the other hand, the Entropy-based elimination retains its pessimism, discarding 26% of the workers.

Table 4: Percentage of Eliminated Workers

Gold Questions (70% accuracy level)	Skill-Adapted Gold Questions	Entropy-based Elimination	Rasch-based Elimination
32% (13 workers)	14% (6 workers)	26% (11 workers)	16.6% (7 workers)

Note that unlike the laboratory-based experiment, no ground-truth is available. Accordingly, to measure the efficiency of the techniques, we evaluate the set of workers who would have been discarded by static gold questions, yet retained by our techniques, namely: the Skill-Adapted Gold Questions and the Rasch-based Elimination. We provided each of those workers with HITS of 10 gold questions that corresponded to their skill level, at which point they provided on average 74% correct answers (i.e. providing only 2 incorrect answers in a HIT), which even beats the initial accuracy level threshold set by static gold questions.

7 Summary & Outlook

In this paper, we addressed the central question of a fair choice of workers that concerns impact sourcing: distinguishing between honest workers, who might just be lacking educational skills, and unethical workers. Indeed, our laboratory study shows how current quality control measures (namely gold questions and reputation-based systems) tend to misjudge and exclude honest and willing workers who may just not have been provided tasks on their individual skill levels. Yet, impact sourcing has to be fair by providing tasks for honest and willing workers that are well within their skill level, while avoiding unethical workers to keep up result quality constraints.

Accordingly, we developed a framework to promote fair judgments of workers. At its heart we deployed the Rasch model, which takes into account both, the level of a worker's skill and the task's difficulty. This aids in distinguishing each worker's success zone. We then presented three advanced, yet practical techniques for detecting irrational patterns within success zones: 1) Skill-adapted gold questions, 2) Entropy-based Elimination and 3) Rasch-based Elimination, with the first and third techniques being optimistic and the second more pessimistic in nature. Both laboratory-based and real world crowd souring experiments attested to the efficiency of the techniques: about 50% of misjudged honest workers can be successfully identified, and subsequently be redirected to skill-fitted tasks.

So far, the Rasch model can be applied on Gold Questions to compute tasks' difficulties, in future work we will expand our framework to majority vote scenarios, too. Due to the redundant nature of majority voting, here interesting financial gains can be realized. Of course, this also needs the computation of task difficulties of non-gold questions, which would allow for a dynamic adaptive task allocation fitting the corresponding workers' skills. Thus, skill ontologies and competence profiles are bound to be of major importance.

References

[1] F. Gino and B. R. Staats, "The microwork solution," *Harvard Business Review*, vol. 90, no. 12. 2012.

[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. on Very Large Data Bases (VLDB)*, pp. 538–549, 2012.

[3] J. Howe, “The Rise of Crowdsourcing,” The *Jounral of North*, vol. 14, no. 14, pp. 1–5, 2006.

[4] D. Zhu and B. Carterette, “An Analysis of Assessor Behavior in Crowdsourced Preference Judgments,” in *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, no. Cse, pp. 17–20, 2010,.

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

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

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

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

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

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

[11] 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

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

[13] 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.

[14] 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 Proceedings of NIPS,* vol. 22, no. 1, pp. 1–9, 2009.

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

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

[17] A. I. Qiang Liu, Mark Steyvers, “Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?,” *In Proceedings of NIPS*, 2013.

[18] 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.

[19] A. Ignjatovic, N. Foo, and C. T. L. C. T. Lee, “An Analytic Approach to

Reputation Ranking of Participants in Online Transactions," *IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 1, 2008.

[20] 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.

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

[22] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer, "An Evaluation of Aggregation Techniques in Crowdsourcing," *Web Information Systems Engineering – WISE* 2013.

[23] J. Wang, P. G. Ipeirotis, and F. Provost, "Managing Crowdsourced Workers," *Winter Conference on Business Intelligence*, 2011.

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

[25] T. G. Bond and C. M. Fox, "Applying the Rasch Model: Fundamental Measurement in the Human Sciences," *Journal of Educational Measuremen*t, Volume 40, Issue 2, pages 185–187, 2003.

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

[27] J. M. Linacre, "Understanding Rasch measurement: estimation methods for Rasch measures," *Jounral of Outcome Measurment*, vol. 3, no. 4, pp. 382–405, 1999.

[28] L. G., "A basis for scaling qualitative data," *Journal of American sociological review*, pp. 139-150. 1944. .

[29] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mobile Computing and Communications Review* , Volume 5 Issue 1, 2001

[30] G. Rasch, "On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements," *Jounal of Danish Yearbook of Philosophy*, vol. 14, 1977.