Using Worker Self-Assessments for Competence-based Pre-Selection in Crowdsourcing Microtasks

UJWAL GADIRAJU, L3S Research Center, Leibniz Universität Hannover
BESNIK FETAHU, L3S Research Center, Leibniz Universität Hannover
RICARDO KAWASE, mobile.de GmbH / eBay Inc.
PATRICK SIEHNDEL, L3S Research Center, Leibniz Universität Hannover
STEFAN DIETZE, L3S Research Center, Leibniz Universität Hannover

Paid crowdsourcing platforms have evolved into remarkable marketplaces where requesters can tap into human intelligence to serve a multitude of purposes, and the workforce can benefit through monetary returns for investing their efforts. In this work, we focus on individual crowd worker competencies. By drawing from self-assessment theories in psychology, we show that crowd workers often lack awareness about their true level of competence. Due to this, although workers intend to maintain a high reputation, they tend to participate in tasks that are beyond their competence. We reveal the diversity of individual worker competencies, and make a case for competence-based pre-selection in crowdsourcing marketplaces. We show the implications of flawed self-assessments on real-world microtasks, and propose a novel worker pre-selection method that considers accuracy of worker self-assessments. We evaluated our method in a sentiment analysis task and observed an improvement in the accuracy by over 15%, when compared to traditional performance-based worker pre-selection. Similarly, our proposed method resulted in an improvement in accuracy of nearly 6% in an image validation task. Our results show that requesters in crowdsourcing platforms can benefit by considering worker self-assessments in addition to their performance for pre-selection.

CCS Concepts: •**Human-centered computing** →**Human computer interaction (HCI);** •**Applied computing** →*Psychology;*

Additional Key Words and Phrases: Crowdsourcing, Pre-selection, Pre-screening, Self-assessment, Microtasks, Performance, Worker Behavior

ACM Reference format:
Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2016. Using Worker Self-Assessments for Competence-based Pre-Selection in Crowdsourcing Microtasks. 1, 1, Article 1 (January 2016), 25 pages.
DOI: 0000001.0000001

This work is supported by the European Commission within the H2020-ICT-2015 Programme (AFEL project, Grant Agreement No. 687916).
Author's addresses: Ujwal Gadiraju, Besnik Fetahu, Patrick Siehndel and Stefan Dietze, Appelstrasse 4, L3S Research Center, Leibniz Universität Hannover, Hannover 30167, Germany; Ricardo Kawase, mobile.de GmbH/eBay Inc., Berlin, Germany.

DOI: 0000001.0000001

1 INTRODUCTION

Researchers and practitioners have actively been both studying and exploiting the crowdsourcing paradigm over the last decade. A recent report regarding the state of *crowdsourcing* in the year 2015 has shed light on the remarkable adoption of crowdsourced solutions to solve a multitude of problems in various industries[1].

Typically in a paid microtask crowdsourcing system, a worker accesses the tasks available and chooses which task(s) to complete. The factors that influence a worker's choice in task selection have been studied in detail in previous works [15, 18]. The self-centric and subjective nature of task selection on a large crowdsourcing platform (such as Amazon's Mechanical Turk[2] or CrowdFlower[3]) is apparent, i.e., it is up to the crowd workers to select a task according to their interests, preference, or expertise. The increasing popularity of crowdsourcing microtasks along with the range of platforms facilitating such efforts, can lead to an overload of choices for a crowd worker. As pointed out by Barry Schwartz in his influential psychology and social theory works, an overload of choices often tends to have detrimental effects on the decision making process of people [39, 40]. The large variety of choices in the tasks that are available for an experienced crowd worker [5] makes it difficult for one to select an appropriate task to complete; workers struggle to find tasks that are most suitable for them.

Prominent marketplaces like Amazon's Mechanical Turk (AMT) or CrowdFlower, that serve as intermediaries to numerous other crowdsourcing channels, gather and accumulate large numbers of diverse tasks. The effort required to search for suitable tasks (in terms of a workers' competencies or interests), or in some cases a lack of alternatives [15], leads to workers settling for less suitable tasks. The quality of the work thus produced eventually decreases. This is supported by the findings of [5], where the authors found that workers most often choose tasks from the first page of the 'recently posted tasks', or the first two pages of 'tasks with most available instances'. More recently, a study of the dynamics of microtasks on AMT by Difallah et al. showed that freshly published tasks have almost ten times higher attractiveness for workers as compared to old tasks [6]. While some workers settle to work on tasks that are not optimally suited to them, some more capable workers may be deprived of an opportunity to work on the tasks they are ideally suited for, due to limitations on the number of participants or individual contributions. Workers often participate in tasks which are beyond their competence and skills, despite their inherent attempt to maintain their reputation. Thus, the overall effectiveness of the crowdsourcing paradigm decreases.

In order to solve the problem of unsuitable workers participating in tasks, pre-selection of workers is the popularly adopted solution [36]. Such pre-screening methods are generally based on the performance of workers on prototypical tasks. If a worker passes a prototypical task or a qualification test, then she can proceed to participate in the actual task. This means that the performance of a worker in a prototypical task is assumed to be an indicator of the competence of a worker. In this work, we draw from self-assessment theories in psychology and organizational behavior in order to show that crowd workers often lack an awareness regarding their competence. We build on these theories which suggest that true competence goes hand-in-hand with the *awareness* of competence, or the lack of it [9, 10]. In contrast to existing methods, we show that by using worker self-assessments as an indicator of competence alongside performance in the pre-screening phase, one can facilitate pre-selection leading to better results in paid crowdsourcing microtasks.

The main contributions of our work stem from (a) investigating whether flawed self-assessments (based on the Dunning-Kruger effect, described in the following section) are prevalent in crowd workers within the microtask crowdsourcing paradigm, and (b) studying the use of self-assessments for worker pre-selection in crowdsourced microtasks. Our contributions are listed below.

[1] http://bit.do/eyeka-crowdsourcing-trend-report

[2] https://www.mturk.com/mturk/

[3] http://www.crowdflower.com/

- By establishing that some crowd workers fall prey to flawed self-assessments, we show that not all workers are aware of their true competence.
- We show that a worker's estimate of her competence in a task is affected by the objective difficulty-level of the task.
- We show that by using rapidly-prototyped self-assessments within the pre-selection process, requesters can ensure that relatively more competent crowd workers participate in their tasks.
- We evaluated our proposed method on a real-world sentiment analysis task and an image validation task, and found an improvement in the quality of results by over 15% and 6% respectively when compared to the existing state-of-the-art pre-selection method.

2 BACKGROUND AND RESEARCH QUESTIONS

2.1 Dunning-Kruger Effect

The Dunning-Kruger effect is a cognitive bias that entails inflated self-assessment and illusionary superiority amongst incompetent individuals [9]. The authors proposed that incompetence in a particular domain reduces the metacognitive ability of individuals to realize it. Skills that encompass competence in a particular domain are often the same skills that are necessary to evaluate competence in that domain. For example, consider the ability to solve a Math problem; the skills required to solve the problem are the same skills that are necessary in order to assess whether the Math problem has been accurately solved. The authors attribute this bias to the metacognitive inability of incompetent individuals. On the other hand, competent individuals tend to underestimate their relative competence due to falsely assuming that tasks that they find easy are also easy for others. The authors thereby show that incompetent individuals cognitively miscalibrate by erroneously assessing oneselves, while competent individuals miscalibrate by erroneously assessing others. In their studies, the authors investigate the self-assessment of individuals over 4 quartiles of their performance distribution. We compute the quartiles such that the top-quartile consists of individuals whose performance score falls in the top-25% of all scores, and the bottom-quartile consists of individuals whose performance score falls in the bottom-25% of all scores, as shown in Figure 1.

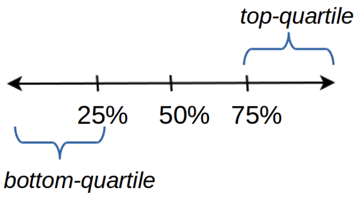


Fig. 1. The Dunning-Kruger Performance Quartiles.

2.2 The Domain of Microtask Crowdsourcing : Motivation

Kruger and Dunning consolidated their findings through 4 studies that addressed a total of 350 Cornell University undergraduate students [27]. In our work, we investigate whether the Dunning-Kruger effect can be observed in the paid microtask crowdsourcing paradigm. The characteristic features of paid microtask crowdsourcing are very different in comparison to the controlled environment where undergraduate students were studied. Firstly, there is a large diversity in the demographics of crowd workers [21, 37]. Secondly, crowd workers have varying motivations to participate in microtask completion, resulting in a wide range of behavior [15, 18]. Thirdly, while the authors rewarded students with credit points for participating in their studies, we provide monetary incentives to crowd workers. It is noteworthy that our study addresses a considerably larger magnitude of participants (over 2,000 crowd workers). Finally, task difficulty for workers can vary across different tasks. In this paper, we will use the following terms to refer to crowd workers with different skills.

Definition 1. *Competent workers* are those crowd workers whose performance in a task lies within the *top*-quartile.

Definition 2. *Least-competent workers* are crowd workers whose performance in a task lies within the *bottom*-quartile.

2.3 Research Questions and Methodology

We address the following research questions in this paper.

RQ#1. Can the Dunning-Kruger effect bear implications on the quality of crowdsourced work?
RQ#2. How are crowd worker self-assessments affected by the inherent level of difficulty in a given task?
RQ#3. Can accurate self-assessments of a crowd worker contribute to realize a stronger indicator of the worker's competence, when compared to performance alone in the pre-screening phase of a given task?

Based on the Dunning-Kruger effect, we adapt the following hypotheses (I, II, and III) to fit the crowdsouring paradigm. We presume that by investigating these hypotheses, we can establish the existence and extent of the Dunning-Kruger effect among crowd workers in paid microtask crowdsourcing platforms.

Hypothesis I. *Least-competent crowd workers overestimate their performance with respect to the competent workers, relative to certain objective criteria.* An example of objective criteria in this context is score in a given test.

Hypothesis II. *Least-competent crowd workers are less capable of identifying competence in themselves or other workers, in comparison to competent workers.*

Hypothesis III. *Least-competent crowd workers are less capable of identifying competence in themselves given the responses of the rest of the crowd, in comparison to competent workers.*

To validate the hypotheses we carry out two studies; in *Study-I* we assess whether crowd workers are aware of their competence, drawing comparison between competent and least-competent workers (addressing Hypothesis I, II). In *Study-II* we investigate whether knowledge about responses of other workers has an effect on the performance of competent and least-competent workers (addressing Hypothesis III).

In studies *III, IV* we evaluate whether considering self-assessments of crowd workers can result in realizing a stronger indicator of their true competence. We propose the pre-selection of workers based on their performance and self-assessments, as opposed to traditional pre-selection based on performance alone. In *Study-III* we consider the task of sentiment analysis, and in *Study-IV* we consider an image validation task, since they are popular examples of real-world crowdsourcing microtasks.

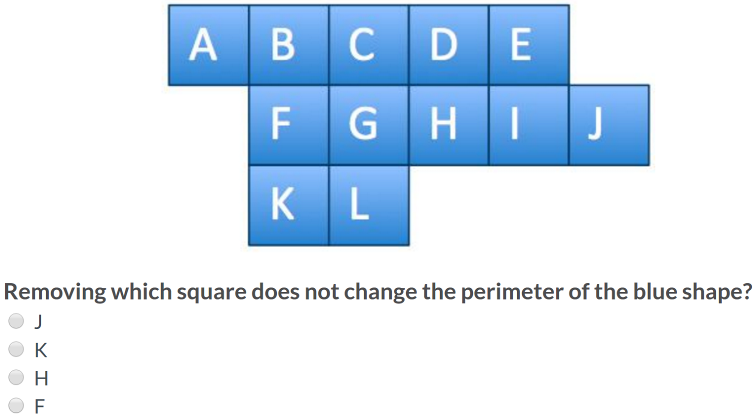
3 STUDY I : SELF-ASSESSMENT OF CROWD WORKERS

Aiming to gather responses from crowd workers and investigate the pre-stated hypotheses (I, II), and to analyze the diversity in competence among crowd workers, we consider the domain of logical reasoning (as in [27]).

3.1 Microtask Design

The task begins with some basic background and demographic questions. It is then followed by 15 questions in the domain of *logical reasoning*. We used logical reasoning questions from *A + Click*[4], where the questions are based on the Common Core Standards[5]. The Common Core is a set of academic standards in Mathematics and English. These learning goals indicate what a student should know and be able to do at the end of each grade.

[4] http://www.aplusclick.com/
[5] http://www.corestandards.org/

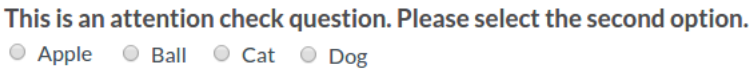


Removing which square does not change the perimeter of the blue shape?

○ J
○ K
○ H
○ F

Fig. 2. An example logical reasoning question from *A+Click* that was administered to crowd workers in the task corresponding to Grade 5.

To assess the varying competencies among crowd workers and the effect of task difficulty, we deployed 8 tasks on CrowdFlower[6] that are designed similarly except for the difficulty level of the logical reasoning questions. We used graded questions from $A + Click$ to administer logical reasoning questions from the level of Grade 5 to Grade 12. An example is presented in Figure 2. Initial empirical tests showed that crowd workers tend to achieve nearly perfect accuracy in logical reasoning tasks that correspond to grades lower than 5. We thereby do not scrutinize grades below 5 further. To separate *trustworthy* workers (TW)[7] from *untrustworthy* workers (UW)[8], we intersperse attention check questions recommended by [16, 30] as shown in Figure 3.



This is an attention check question. Please select the second option.

○ Apple ○ Ball ○ Cat ○ Dog

Fig. 3. Attention check questions to identify *untrustworthy* workers.

At the end of the logical reasoning questions, workers are requested to answer questions in relation to their performance, corresponding to the following aspects.

- **Perceived test score.** Number of questions that the workers believe to have answered correctly. The corresponding question was phrased as follows – *How many questions do you think you answered correctly?* (answer range: 0-15).
- **Perceived test score of others.** Number of questions on average, that workers think others participating in the task will have answered correctly. The corresponding question was phrased as follows – *On average, how many questions do you think the other workers completing this task will answer correctly?* (answer range: 0-15).
- **Perceived ability.** The expected percentile ranking of the workers. The corresponding question was phrased as follows – *At what percentile ranking (1-100) do you expect to be, with respect to all the workers who will perform this task? '1' indicates the very bottom, '50' indicates exactly average, and '100' indicates the very top* (answer range: 1-100).

[6] http://www.crowdflower.com/

[7] Workers who correctly answer all 3 attention check questions embedded in the task.

[8] Workers who incorrectly answer at least 1 of the 3 attention check questions embedded in the task.

Finally, in order to analyze aspects pertaining to real-world tasks, workers were asked to provide as many tags as possible for two different pictures. Tagging images is a popular type of crowdsourced task. Prior research has shown that having verifiable questions such as tags is a recommended way to design tasks and assess crowdsourced results [22].

The order in which different questions were asked did not have an impact on any of the results reported in our work. We thereby do not mention it further. We paid each worker according to a fixed hourly wage of 7.5 USD. In each of the 8 tasks, corresponding to the 8 different graded levels of competencies, we gathered 250 responses from independent workers, resulting in a total of 2,000 crowd workers overall.

3.2 Trustworthiness of Workers

From the responses gathered through the 8 tasks, we first separated trustworthy workers (TW) from untrustworthy workers (UW). Table 1 shows the number of TW out of the 250 workers that participated in total, in each grade. On average each grade has around 216 TW participants.

Table 1. Distribution of Trustworthy Workers (TW) across the graded microtasks.

Grade	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**	**Avg. G5-G12**
#TW	228	216	226	207	207	215	214	219	216.5
TW (in %)	91.2	86.4	90.4	82.8	82.8	86	85.6	87.6	86.6

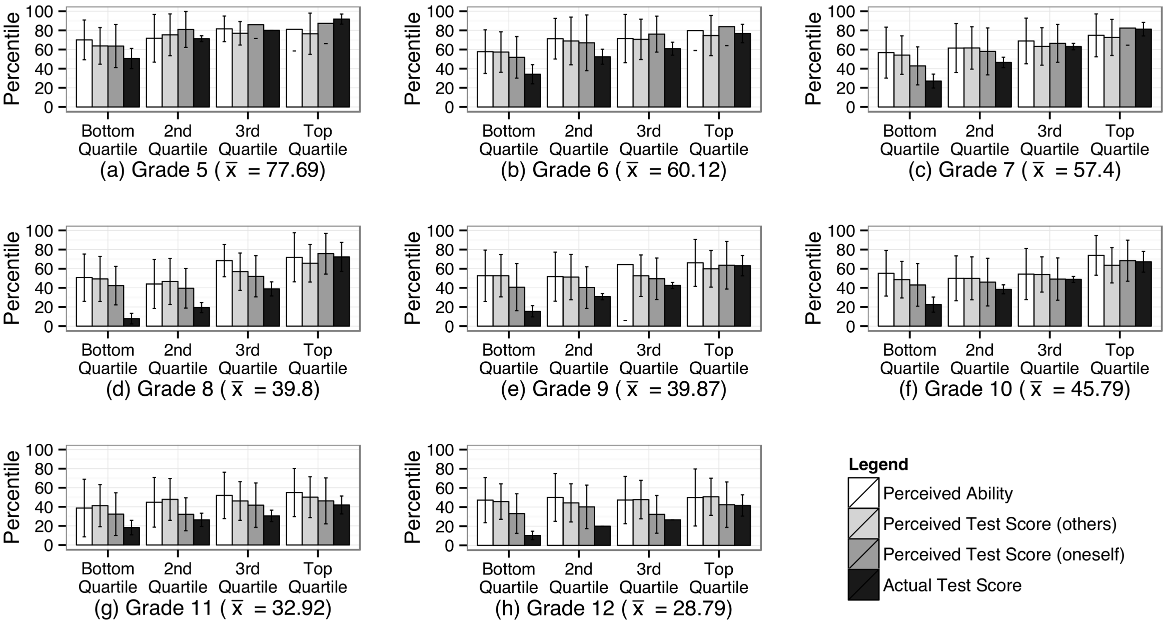
To establish a correlation between the country of origin and the performance of a worker, several experiments that consider aspects such as the time of task deployment, batch size, channels used, and so forth are needed. Addressing the implications of cultural differences in task performance [33] is beyond the scope of this work. Note that we do not consider the UW in the rest of our study and analysis.

4 RESULTS: SELF-ASSESSMENT OF CROWD WORKERS

4.1 Perceived Test Score and Ability of Oneself

We analyzed the responses of each worker for the questions pertaining to *perceived test score* and *perceived ability*. Our findings are presented in the Figure 4. We observe that through all the grades (G5-G12), the least-competent workers (i.e. bottom-quartile workers) significantly overestimate their ability and raw test scores. We find that Figure 4(a) represents a perfect scenario of the dual fallacy resulting in the self-assessments observed by [27]. The least-competent workers overestimate their ability (by nearly 20 percentile points) and performance (by around 13 percentile points). Hence, we observe that least-competent crowd workers cognitively miscalibrate by erroneously assessing themselves, while competent crowd workers miscalibrate by erroneously assessing others (they underestimate their ability by 10 percentile points and performance by nearly 4 percentile points).

With the increase in grade levels (from G5 through G12), we note that least-competent workers depict an increase in the degree of overestimation in the assessment of their ability and performance (perceived test score). A novel finding through our work pertains to that of the competent workers. We note that with an increase in grade levels, competent workers also tend to gradually shift towards overestimation of their ability and performance. We attribute this to the increasing grade levels which potentially go beyond their competence at some point. However, it is clear that least-competent crowd workers indeed overestimate their ability and performance by several percentile points (*M=30.18, SD=9.53*) in comparison to the competent crowd workers (*M=-3.73, SD=7.79*) across all grades (*t(13)=4.22, p<.001*). We found a very large effect size; *Cohen's d = 3*. Thus, we found support for **Hypothesis-I**.



(a) Grade 5 ($\overline{x}$ = 77.69) (b) Grade 6 ($\overline{x}$ = 60.12) (c) Grade 7 ($\overline{x}$ = 57.4)

(d) Grade 8 ($\overline{x}$ = 39.8) (e) Grade 9 ($\overline{x}$ = 39.87) (f) Grade 10 ($\overline{x}$ = 45.79)

(g) Grade 11 ($\overline{x}$ = 32.92) (h) Grade 12 ($\overline{x}$ = 28.79)

Fig. 4. Perceived test scores and perceived ability of workers across the graded microtasks. Quartiles are presented on the x-axis, percentile on the y-axis, and '$\bar{x}$' is the mean performance of workers in the corresponding grade.

4.2 Perceived Test Score of Others

From the plots in Figure 4, we clearly observe that least-competent crowd workers greatly miscalibrate their assessment of others in terms of the raw test score (i.e., the number of questions answered correctly by others on average). For instance, consider the Figure 4(a) corresponding to G5. Here, least-competent crowd workers placed the average performance of other workers in the 63rd percentile, while the actual mean performance was 77. Competent crowd workers on the other hand fractionally overestimated the average performance of others and placed it in the 78th percentile.

Interestingly however, we note that with the increasing grade levels from G6 through G12, both competent and least-competent workers overestimate the average performance of other workers. Moreover, we find that the degree of *miscalibration* (i.e., the difference between the actual score of a worker and the worker's perceived test score) is more prominent with respect to competent workers. While the actual mean performance was in the 39th percentile on average across all grades, the least-competent workers overestimate the performance of others by around 14 percentile points, and the competent workers overestimate the performance of others by around 25 percentile points. We believe that due to the increasing difficulty inherent to progressive grade levels, competent workers tend to further miscalibrate their relative competence and least-competent workers start recalibrating their relative competence in the accurate direction. Due to the fact that competent workers tend to wrongly believe that their peers are of relatively good competence, they overestimate the performance of others to a greater extent in the higher grades. We thereby found that across all grades competent workers overestimate the performance of others by more percentile points (*M=17.16, SD=7.71*) than incompetent workers (*M=9.41, SD=5.65*), *t(13)=3.01, p<.005*, with a large effect size; *Cohen's d = 1.15*. Thus, we did not find full support for **Hypothesis-II**, stating that 'least-competent crowd workers are less capable of identifying competence in themselves or other workers in comparison to competent workers'.

5 STUDY II : SELF-ASSESSMENT IN THE PRESENCE OF OTHERS' ANSWERS

To assess whether least-competent workers are capable of identifying their true level of competence given the performance of the rest of the crowd (hypothesis III), we deployed a second set of tasks on CrowdFlower.

5.1 Microtask Design

Since we aim to draw a comparison between the competent and least-competent workers alone, we contacted the top and bottom-quartile workers from our priorly completed graded tasks (in Study I) via e-mail and requested them to participate in the subsequent task for each corresponding grade. Over 70% of the top and 60% bottom-quartile workers participated in these tasks over two weeks from deployment. To make valid comparisons across the different grades, we considered the first 60% in each of the top and bottom-quartile workers that participated. These tasks were identical to the initial 8 graded tasks that were deployed (including the incentives offered), with one exception. In this case, we show the overall answer distributions (see Figure 5) provided by all workers in the initial round of tasks (in a bar graph) alongside each question in the set of 15 logical reasoning questions.

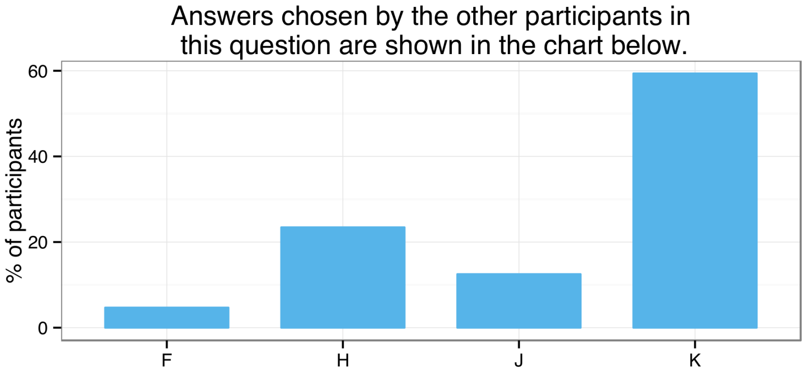


Fig. 5. The overall answer distribution corresponding to a sample question from Grade 5 (see Figure 2), that is shown alongside the question in *Study-II*.

5.2 Results (Study II)

Figure 6 presents our findings. We observe that the overall mean performance for each grade improves in comparison to the first set of tasks. This is expected since the workers participate in the same task for the second time. In addition, the workers are aided by the distribution of answers for each question, since they can go with the majority in case they are unsure about certain answers. Our observations are further validated by the completion time of workers in the top and bottom quartiles. In case of top-quartile workers, showing the distribution of answers for each question results in significant reduction in completion time (*M=4.14, SD=1.88*), *t(14)=4.14, p<.001, Cohen's d = 1.34* with a reduction of 4.1 minutes on average across all grades. However, this is not the case for the bottom-quartile workers (*M=1.39, SD=2.21*), where the difference in completion time is not significant, and the reduction in completion time is only 1.39 minutes on average. We also found that the poorly

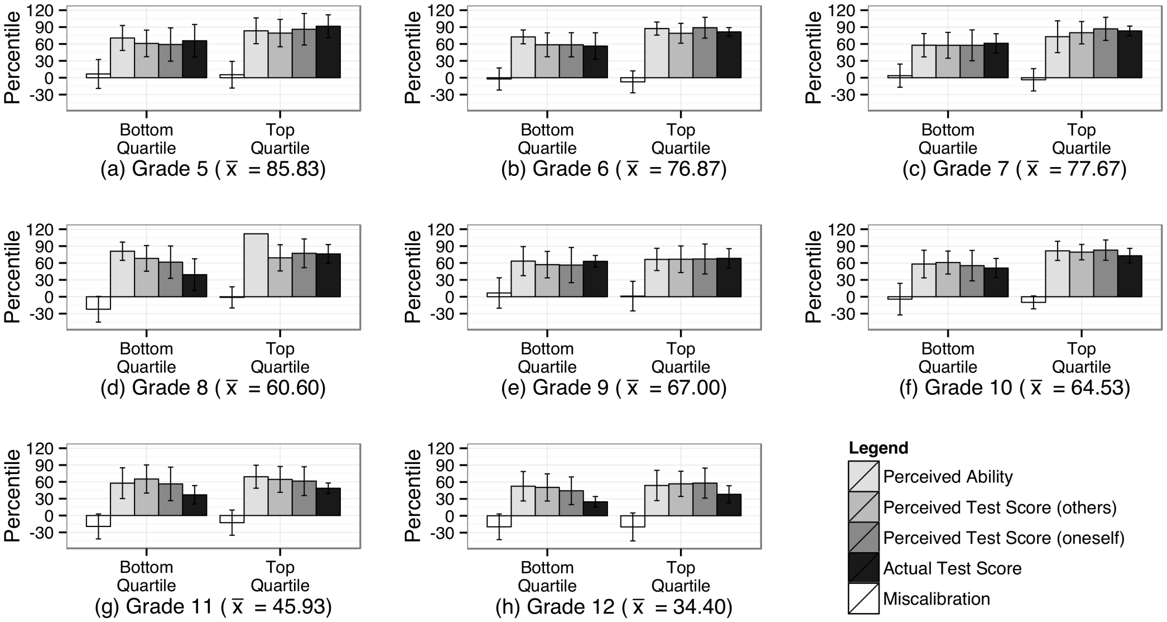


Fig. 6. Perceived test scores and perceived ability of *competent* and *least-competent* workers across the graded microtasks (G5–G12), in the presence of overall answer distributions for each question. Quartiles are represented on the x-axis, and the y-axis represents percentages of the corresponding attribute. '$\bar{x}$' is the mean performance of workers in the grade.

performing workers in the first round of tasks take much less time for completing the task (as shown in Figure 15 in the Appendix). Hence, the room for reducing task completion time further is minimal.

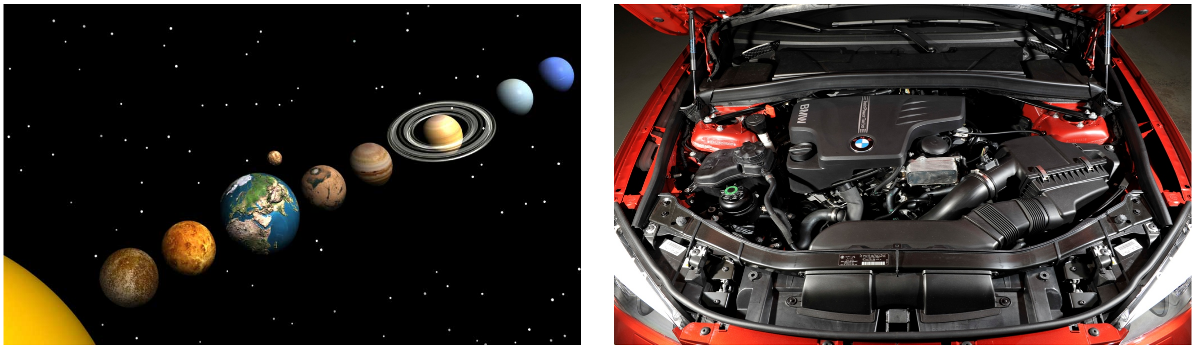
In grades G6 and G10, the competent workers depicted a greater degree of miscalibrated self-assessment when compared to the least-competent workers. We thereby note that the *miscalibration* (i.e., the difference between the actual score of a worker and the worker's perceived test score) of least-competent workers is more pronounced, although it is inconsistent. Therefore, we find only partial support for **Hypothesis III**.

6 IMPLICATIONS ON REAL-WORLD MICROTASKS

Through our findings from Study I and Study II, we can conclude that the Dunning-Kruger effect can be observed in the crowd, subject to the task difficulty at hand. To understand what this difference in competence between top-quartile workers and bottom-quartile workers means in terms of their performance in a real-world microtask, we investigate the tagging task that workers completed at the end of each of the graded tasks in Study I.

To observe the implications of worker competence on a traditional crowdsourcing task like *tagging* (a popular example of content creation tasks [15]), we analyzed the tags received from least-competent (bottom-quartile) and competent (top-quartile) workers in Study I. Workers were asked to provide as many tags as possible, corresponding to two pictures presented as shown in Figure 7. We first processed the responses from crowd workers, so as to ignore meaningless phrases and gibberish tags. We evaluate tags with respect to *quality* (i.e., the reliability of a tag[9]) and *quantity* (i.e., the number of tags).

[9] A tag that is mentioned by at least 10 distinct workers is defined as a *reliable tag.*



(a) Picture 1 (Pic#1) – the solar system.

(b) Picture 2 (Pic#2) – the engine of a car.

Fig. 7. Pictures corresponding to the tagging task that workers were asked to complete at the end of *Study-I*.

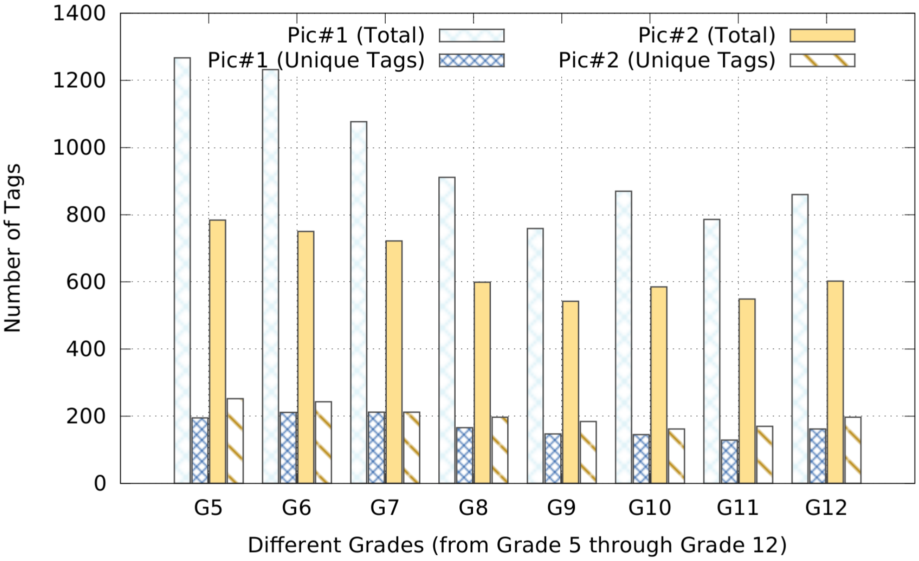


Fig. 8. Distribution of tags contributed by workers in each of the grades (G5-G12).

Figure 8 shows that the total number of tags and unique tags provided by workers decreased gradually with the increase in grade level (adjusted for worker distribution across grades, see Table 1). This implies that due to the increasing difficulty with progressive grades, workers exert relatively less effort in providing tags. This is in accordance with findings from prior works that have explored the effect of one microtask on another, and between those with varying difficulty levels [4, 35]. Corresponding to Pic#1, there were a total of 1,267 tags with 195 unique tags for Grade 5, when compared to 860 tags with 162 unique tags for Grade 12. In case of Pic#2, there were a total of 784 tags with 252 unique tags for Grade 5. This decreased to a total of 692 tags with 197 unique tags for Grade 12.

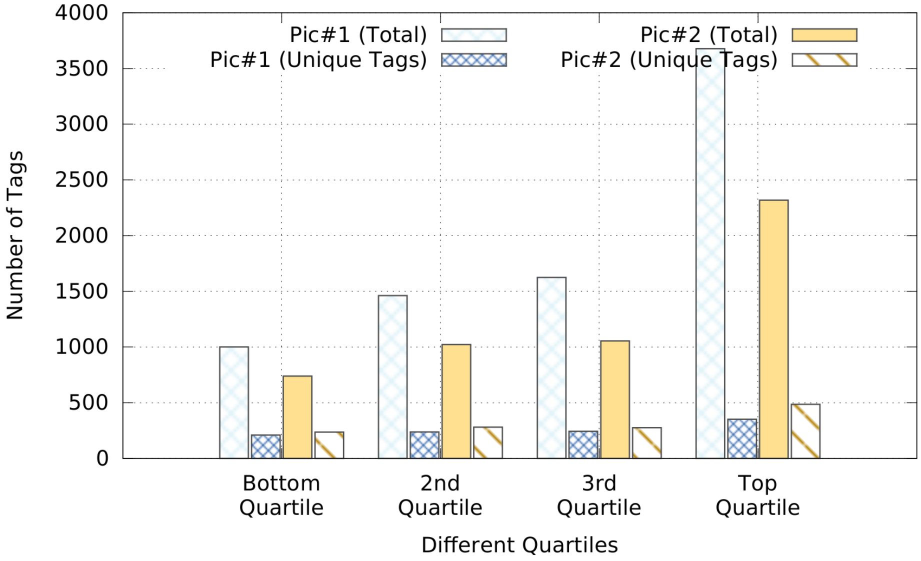


Fig. 9. Distribution of tags from all grades (G5-G12) across the quartiles.

We did not find a significant difference in the *quantity* of *reliable tags* across the different grades (G5-G12). On average, for each grade there were around 18 reliable tags corresponding to Pic#1, and nearly 7 corresponding to Pic#2. Figure 9 presents the distribution of tags from all grades with respect to the performance quartile. We found that competent workers provided more distinct reliable tags (31 for Pic#1, 25 for Pic#2) than least-competent workers (18 for Pic#1, 8 for Pic#2). These differences in number of reliable tags produced by the competent workers (*M=18.75, SD=3.62*) and least-competent workers (*M=3.75, SD=2.23*) across the grades are found to be statistically significant, $t(11)=4.43$, $p < 0.01$. Our findings suggest that competent (top-quartile) workers provide more reliable tags, with a higher diversity, when compared to least-competent (bottom-quartile) workers.

Operationalizing Worker Self-Assessments

From our findings in Study I and II it is evident that not all crowd workers are adept at making accurate self-assessments; competent workers are relatively better at doing so. This is further reinforced by our findings in the tagging task, where we observe that top-quartile workers produce tags with both a higher quality as well and quantity. Based on this understanding, we propose that it can be beneficial to operationalize worker self-assessments as an indicator of worker competence and therefore performance. To do so, we propose to use accuracy of worker self-assessments in the pre-screening tasks in addition to their actual performance in the pre-screening tasks to select workers. Thus, the only additional requirement in our proposed method is a self-assessment question at the end of the pre-screening tasks, making it straightforward to implement. Figure 10 illustrates the traditional pre-screening method in comparison to our proposed self-assessment based pre-screening approach.

7 STUDY III: EVALUATION IN SENTIMENT ANALYSIS TASK

From our findings in Study I and II we note that some crowd workers (bottom-quartile) exhibit inflated self-assessments. We also found that the top-quartile workers produce significantly better quality of work, as observed

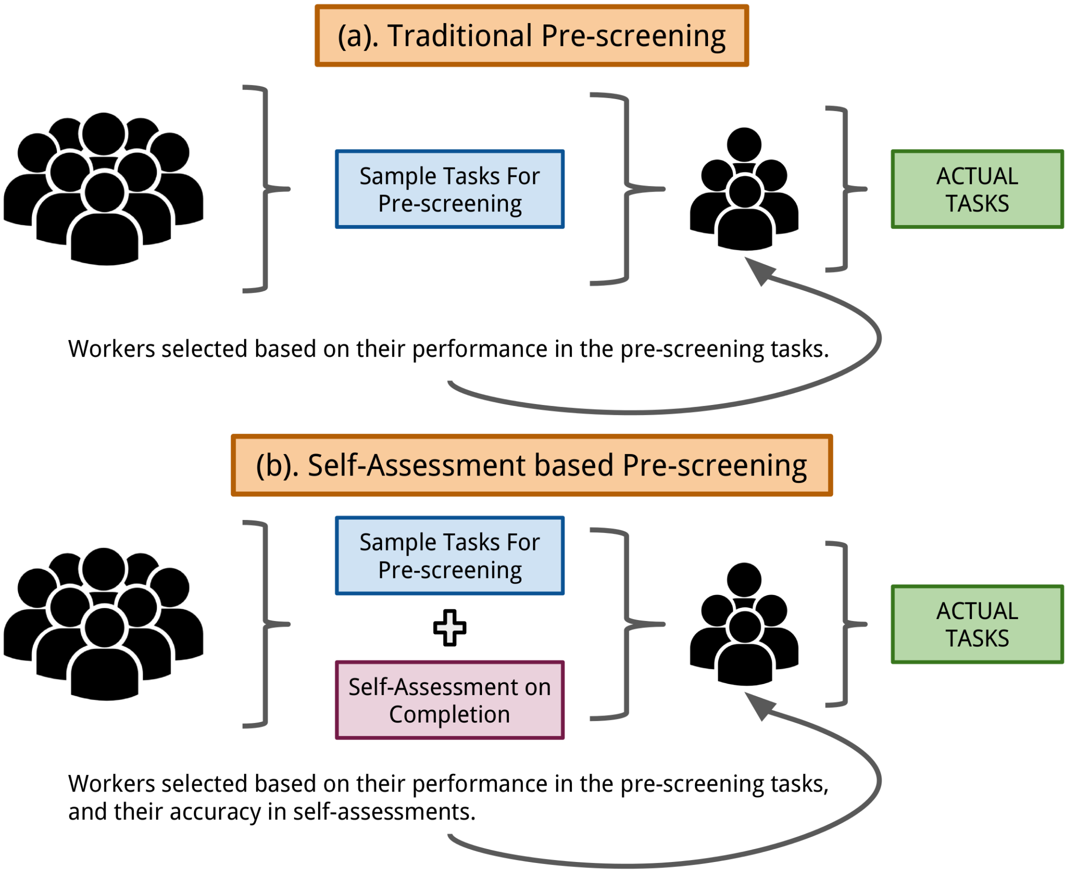

(a). Traditional Pre-screening
Sample Tasks For Pre-screening
ACTUAL TASKS
Workers selected based on their performance in the pre-screening tasks.
(b). Self-Assessment based Pre-screening
Sample Tasks For Pre-screening
Self-Assessment on Completion
ACTUAL TASKS
Workers selected based on their performance in the pre-screening tasks, and their accuracy in self-assessments.

Fig. 10. Comparison between (a) the traditional pre-screening method based on worker performance in pre-screening tasks, and (b) the self-assessment based pre-screening method which considers worker performance in the pre-screening tasks as well as their accuracy in self-assessments.

in the abridged tagging task of Study I. In Study III, we seek to answer whether we can operationalize the ability of workers to accurately self-assess their performance in a real-world microtask, in order to pre-select a more suitable crowd with respect to the task. Can worker self-assessments be used as a means to provide a stronger indicator of worker competence (**#RQ3**)?

We evaluated our proposed method of using worker self-assessments as a basis for pre-screening crowd workers, as opposed to traditional pre-screening that is purely based on the performance of workers. We considered a popular crowdsourcing task; *sentiment analysis* [15]. In this task composed of 30 units, crowd workers are asked to read a tweet in each unit and classify the projected sentiment as either `positive`, `negative` or `neutral`. For this purpose we use the dataset introduced by [14], that consists of expert-classified tweets, thereby providing our ground truth.

7.1 Method I : Self-Assessment Based Pre-screening

We prototyped a 5-unit task for the sentiment analysis, consisting of tweets different from those in the actual 30 units considered for the evaluation task. On completing these 5 units, workers are asked the question, '*How many questions do you think you answered correctly?*'. We consider a worker to have passed this screening task, if the worker accurately predicts her score while the actual score is > 3, or if the worker miscalibrates her prediction by one point while her actual score is > 3 (i.e., $miscalibration = 0$ or 1). The intuition behind using a threshold of '3' is due to our aim to replicate a realistic pre-selection scenario. CrowdFlower suggests a minimum accuracy of 70% by default[10] for the traditional pre-screening method (which is actual score > 3 in our case). We deployed this task on CrowdFlower and gathered responses from 300 workers by offering a compensation of 2 USD cents. We found that only 110 out of 300 workers passed the threshold of actual score > 3/5. Of these 70 workers passed the self-assessment accuracy criteria and thereby passed the pre-screening. Next, we deployed the actual evaluation task consisting of 30 units to these 70 workers alone[11] by using their e-mail IDs. We offered a reward of 5 USD cents to workers. Within a span of 1 week, 50 of the 70 workers completed the task.

7.2 Method II : Traditional Pre-Screening

One week later, we deployed an identical task consisting of the same 30 units on CrowdFlower. There was no overlap in the pool of workers across the two tasks. Hence, the observed results are not due to ordering effects. We used the same 5 units in the traditional pre-screening process as in the case presented above, and only those workers who answered > 3 units correctly were allowed to participate in the actual task. We gathered responses from 50 distinct workers, and these workers were also paid a compensation of 5 USD cents (to match the incentive offered and number of collected judgments in the self-assessment based pre-screening method.

7.3 Results

We evaluated the two different methods based on the following two aspects: accuracy of the pre-selected workers in the tasks following the screening, and their task completion time. We found that the self-assessment based pre-screening method (green dots in Figure 11) resulted in workers who performed with an accuracy of nearly 94% on average, with an inter-annotator agreement of 0.95 (computed by pairwise percent agreement (PPA)). The traditional pre-screening method (presented in Figure 11 in the red color) resulted in workers who performed with an average accuracy of around 78%, with an inter-annotator agreement of 0.83 (computed by PPA).

We found that the difference in the resulting worker performances between using the self-assessment based pre-screening method (M=27.95, SD=1.79) and the traditional pre-screening method (M=23.63, SD=6.23) was statistically significant *t(95)=3.40, p < 0.01*, with a large effect size; *Cohen's d = .94*. We did not find a significant difference in the task completion time of workers resulting from the two different methods of pre-screening.

It is important to note that in the self-assessment based pre-screening method, the average actual scores of workers on the qualification test was 4.4/5 and that of workers in the traditional pre-screening method was 4.3/5, without a significant difference. This shows that the observed improvement is due to the consideration of worker self-assessments, and not simply a result of selecting workers who performed better in the pre-screening phase. We highlight that there may be a confound in having workers wait, then self-select to return and complete the actual evaluation task in the self-assessment based pre-screening method. Such workers may be more diligent than workers in the traditional pre-screening method, who immediately began the actual evaluation task. However, due to the number of workers in the pool, the significant differences and the large effect size observed, we believe this does not risk the overall result and does not pose a threat to its validity.

[10] CrowdFlower's guide to test questions and quality control on: https://success.crowdflower.com/hc/en-us

[11] CrowdFlower provides support for this via the *internal workforce*, https://success.crowdflower.com/hc/en-us/articles/202703355-Contributors-CrowdFlower-s-Internal-Channel.

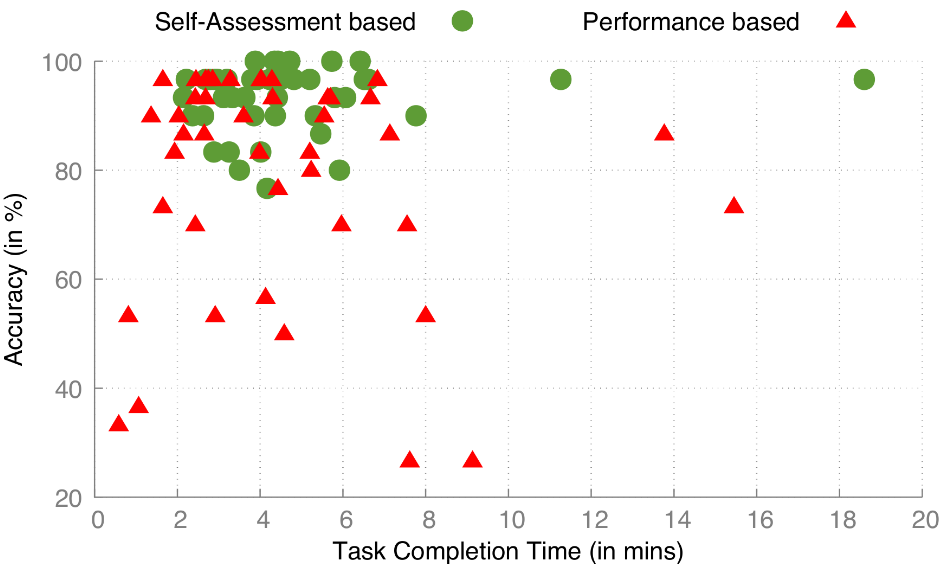


Fig. 11. Performance of workers acquired by the proposed *Self-Assessment based Pre-Screening* and by traditional *Performance based Pre-Screening*

From these results, we observe that pre-screening crowd workers based on their self-assessments provides a better reflection of their actual competence, leading to an improved quality of results. We note an improvement of over 15% in accuracy and 12% in agreement between workers by using self-assessment based pre-screening of workers in a sentiment analysis task. Thus, we can conclude that operationalizing self-assessments of workers in a given task in conjuction to their performance in the task, can serve as a stronger indicator of worker competence than relying on worker performance alone.

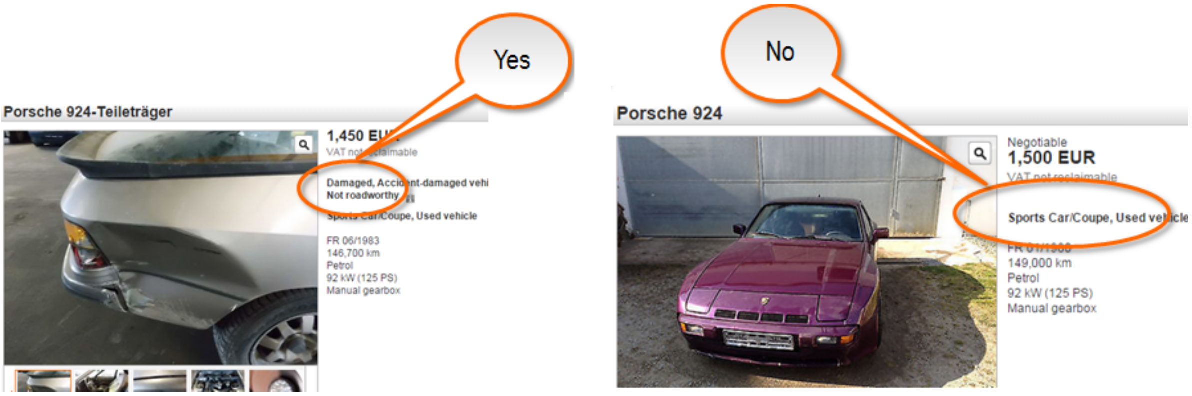
8 STUDY IV: EVALUATION IN VERIFICATION AND VALIDATION TASK

In Study III, we operationalized worker self-assessments in a sentiment analysis task and improved the pre-selection of crowd workers. In Study IV, similar to the sentiment analysis task described in the previous section, we considered an additional real-word task of image validation. Our aim is to verify whether our proposed approach would yield similarly improved results in another type of task, due to the effectiveness of our proposed worker pre-selection method.

In this task composed of 13 units in total, crowd workers were asked to analyze the pictures in online automobile ads to spot mismatched information. To publish an online ad, sellers need to textually describe the state of the vehicle (damaged or not) and its mileage. Sellers commonly omit damage-related information from the description or claim a lower mileage in order to achieve a better placement in the search results (see Figure 12). In many cases this information is evident in the pictures. While this cannot be easily detected by automated algorithms, it is a rather simple task for humans.

8.1 Task Setup

We manually found and annotated a total of 13 vehicle ads[12] which served as groundtruth for the task. Each ad corresponds to one unit where workers are asked to answer three multiple choice questions: (i) Is the car marked as damaged? (ii) Can you identify that the car has a visible damage or functional problems based on the pictures? (iii) Is the mileage information consistent with the picture? We took care to find distinct ads that produced an even distribution of the options corresponding to each question. The units were randomized and after answering 3 units (total of 9 questions), workers were asked to assess their performance on the 9 questions. With an aim to compare self-assessment based pre-screening with performance based pre-screening, all workers were allowed to continue onto 10 more units. Each worker was rewarded with 5 USD cents on successful task completion. We deployed this task on CrowdFlower and collected responses from 100 distinct workers.



(a) Seller declared visible damage in the description of the advertisement.

(b) Seller omitted visible damage-related details from the description of the advertisement.

Fig. 12. Example automobile ads from the online marketplace `mobile.de` that either (a) declare damages in the vehicle description, or (b) omit damage-related information.

8.2 Results

Traditional Pre-screening: Similiar to the previous sentiment analysis task, the traditional pre-screening method is characterized by a performance threshold of 70% in the pre-screening phase. Thus, we filtered out workers (36 in total) who did not achieve a minimum of 70% accuracy in the first 3 units (9 questions). In the 10 units that followed, comprising the actual task, this group of workers (*N=64*) achieved an average accuracy of 84.05% *(M=84.05, SD=10.35*), with an inter-annotator agreement of 0.81 using pairwise percent agreement (PPA). **Self-Assessment Based Pre-screening**: In case of the proposed self-assessment based pre-screening approach, we consider the accuracy of worker self-assessments in addition to the 70% accuracy threshold in the pre-screening phase. Here again, we tolerate an error of 1 point in the workers self-assessments (i.e., $miscalibration$ = 0 or 1). Workers who passed this pre-screening phase (*N=49*), performed with an accuracy of 89.6% ((*M=89.6, SD=6.6*) in 10 units that followed, comprising the actual task. In this case, the inter-annotator agreement was found to be 0.9 (PPA).

[12] We used publicly available ads from the online marketplace http://www.mobile.de/

To summarize, we found that 64 of the 100 workers passed 70% accuracy threshold. Of these, 49 workers passed the self-assessment accuracy criteria and thereby passed the pre-screening. The self-assessment based pre-screening approach resulted in an improvement in accuracy of nearly 6%, and an increase in the inter-annotator agreement between workers by 8% in comparison to the traditional pre-screening method. The difference in worker accuracy between the traditional and the self-assessment based pre-screening methods was found to be statistically significant with a moderately large effect size; *t(112)=2.60, p<.01*, Hedge's *g = .62*. Once again, we noted that the difference in performance in the pre-screening phase (3 units, 9 questions) across the two groups of workers was not statistically significant, indicating that the improvement in the accuracy of workers using our proposed approach is due to the consideration of accuracy of workers' self-assessments. We also did not find a significant difference in the task completion time of workers selected using the different methods.

9 DISCUSSION, CAVEATS AND LIMITATIONS

9.1 Self-Assessments for Competence-based Pre-Selection

Through our experimental findings and evaluation, we observe that using worker self-assessments for competence-based pre-selection can provide a stronger indicator of worker competence and potential performance to requesters. Since workers need to answer only one additional question with regard to estimating their performance, the time overhead in comparison to performance-based pre-selection is negligible. Due to the same reason, rapidly protyping a self-assessment based pre-selection phase requires relatively the same effort from a requester's point of view. Moreover, since there is an improvement in the quality of the results produced, requesters can improve their costs-benefit ratio with respect to a given task. Requesters can also adjust the passing threshold in the pre-selection process to suit their needs. However, this approach may entail a loss in workforce due to more effective pre-selection and thereby increase the overall task completion time. In Study III, around 37% of the 300 workers passed the traditional pre-screening method, while around 24% of the workers passed the self-assessment based pre-screening method. Similarly in Study IV, 64% of the workers passed the traditional pre-screening method and 49% of the workers passed the self-assessment based pre-screening method. On average across the two studies, we note a loss in workforce of less than 14% resulting from the self-assessment based pre-screening method in comparison to the traditional method. Due to the abundance of crowd workers and in the interest of significantly improved results, we believe our proposed approach will lead to meaningful trade-offs. It is important to note that other quality control measures can be easily used in addition to the self-assessment based pre-selection method to further improve the quality of the crowdsourced work.

From our results in Study III and Study IV, we note that the proposed approach yields better results in comparison to traditional pre-screening methods across the two different types of tasks considered. We note that the self-assessments based pre-screening method results in a relatively larger improvement in the sentiment analysis task (considered in Study III) than in the image validation task (considered in Study IV). While this reflects on the generalizability of the proposed approach across task types, the results also indicate that the method can be effective to varying degrees. We reason that this difference is due to the inherent difficulty levels of the task types considered. Further experiments are required to gauge the impact of the proposed approach under the interaction of different task types and task difficulty.

Pre-selection of workers according to our proposed self-assessments based pre-screening approach can mean that workers in such stystems may not get to work on tasks that go beyond their competence. This can be a limitation since challenging tasks can be more interesting and play a developmental role for workers. The resulting potential power imbalance between workers and requesters in terms of using self-assessments for pre-selection, can be overcome by using the self-assessments of workers to also raise their self-awareness, thereby playing a constructive role in supporting the growth [2] of workers and developing crowd work. We will explore the use of self-assessments to increase worker self-awareness in future work.

9.2 Worker Competence Transferability

In our experimental results in the abridged *tagging* task, we observed the implications of competent and least-competent workers on the quality, reliability and the diversity of the tags produced. In this case, we assessed the competence of workers based on the *logical reasoning* task and applied the resulting characterization to the tagging task. By doing so we found that competent workers exhibit a better performance. However, such transferability of worker competence from one type of task to another needs to be studied further. While we cannot assume the universal transferability of a worker's competence that is assessed in one domain alone, an understanding of transferable domains will reduce further costs (in terms of time and money) that are incurred through pre-selection processes. Our proposed approach is to rapidly prototype a given task and use worker self-assessments to assess worker competence in a pre-selection phase. Due to this reason, we carried out further evaluations of worker self-assessment based competence estimation in a sentiment analysis task, and an image validation task.

9.3 Training Crowd Workers to Increase Competence

In their studies, Kruger and Dunning also studied the effect of training less-competent individuals [27]. The authors found that through systemic feedback and training, less-competent individuals can progress towards higher competence, leading them to become more self-aware. However, the impact of learning or training on individuals' self-assessments has attracted several debates on both sides ([38], [34]). While Schosser et al. found no evidence of learning that leads to consequent improvement in performance of incompetent individuals [38], Miller and Geraci cite contrasting evidence through their experiments [34].

Recent works have studied the impact of providing feedback and training workers in crowdsourcing microtasks [14, 29]. Through a series of empirical experiments on different types of microtasks, Gadiraju et al. have shown that the performance of workers can be improved by providing training. In the context of our work in this paper, the findings of Gadiraju et al. [13, 14] can be extrapolated to reason that training least-competent workers can help them improve their competence, and thereby improve the calibration of their self-assessments. However, further scrutiny is required in order to understand the impact of training on crowd workers' self-assessments and competence.

9.4 Other Considerations

It is important to explore whether there are cross-cultural differences in how the Dunning-Kruger effect manifests, that can further dictate the use of self-assessments for pre-selection in tasks. For example, does the perception of their own performance vary across worker groups having different ethnicity? We conducted a one-way between workers ANOVA to compare the effect of *ethnicity* of workers on the perception of their own performance across all grades in 7 ethinicity-group conditions (African American, American Indian, Asian, Hispanic, Pacific Islander, White, Other) as indicated by workers in Study I. We found that there was no significant effect of ethnicity of workers on the perception of their performance at the $p < .5$ level for the 7 ethnicity-group conditions [*F(6, 1725)=0.4229, p=0.86*]. Post-hoc comparisons using the Tukey HSD test confirmed that there was no significant difference in perception of the workers' own performance between any two of the different ethnicity groups.

To give workers a fair chance to participate in a task while using self-assessments as pre-screening method, an important caveat is to ensure that the workers are aware that the selection is based on both, their performance and the accuracy of their self-assessment. Otherwise, workers may inflate their self-assessment with the belief that a higher assessment would lead to their participation in the task. Isolating workers who miscalibrate their self-assessments due to such inflation is beyond the scope of our work. Nevertheless, it is noteworthy that our proposed approach for worker pre-selection is effective in yielding improved results.

10 RELATED LITERATURE

10.1 Self-Assessment

Apart from the priorly discussed work of Kruger and Dunning [27], there have been several other noteworthy works in the realm of individual self-assessment. Research works have shown that people provide inflated self-evaluations on performance in a number of different real world settings. Dunning et al. showed and discussed the implications of such flawed self-assessments on health, educational settings and the general workplace [10].

Kulkarni et al. showed that in an online course addressing a large number of students (MOOC), the students graded their work 7% higher than those assigned by the staff on average [28]. Other existing data from experiments reinforce the mistaken self-evaluation of performance [11, 12]. These works show that incompetent individuals are worse at assessing the quality of performance and often tend to think that they outperform the majority, while in fact they belong to the lower rungs of the performance quartile. Complementing these existing works on self-assessment, in our work we aim to understand whether the flawed self-assessment theories hold among crowd workers in the crowdsourcing paradigm. In contrast to these studies that are largely based on self-selected groups of individuals leading to potential selection bias, we use the crowd as a source for a diverse landscape of individuals with respect to their demographics, skills and competence.

Despite a considerable number of works that assert the findings from the Dunning-Kruger effect, the underlying reasons that dictate the dual-curse resulting in the miscalibrated self-assessment have been widely contested [3, 25, 26]. Several researchers have provided alternative accounts for the Dunning-Kruger effect, alluding it to regression to the mean and the above-average effect. These accounts have in turn resulted in rigorous theoretical responses and empirical refutations [12], and are out of the scope of our work in this paper.

In closely related work that proposes the use of self-assessments to improve crowd work, Dow et al. showed that self-assessments allowed workers to improve over time in a task involving writing consumer reviews of products they owned [7]. The authors of this work proposed the use of self-assessments to yield better work quality by promoting self-reflection and learning. In contrast, we propose to consider the accuracy of worker self-assessments alongside their task accuracy in a pre-selection phase as an indicator of their true competence and potential performance. Thus, we develop a distinct and novel approach by directly leveraging self-assessments as a worker filtering mechanism, rather than aiming to improve work through self-review.

10.2 Competence of Crowd Workers

The crux of prior research works in the realm of characterizing crowd workers has mainly focused on ensuring reliability of workers, and presenting a means to the requester to pre-select prospective workers [23]. In this regard, researchers have suggested the use of pre-screening methods and qualification tests [19], trust models to predict the probability of reliable responses [41], hidden gold standard questions [36], and the use of metrics that quantify acceptability of responses from the crowd [16]. In this paper, we propose a novel method for the pre-selection of workers, that outperforms traditional performance based pre-screening methods.

Kazai et al. [20] used behavioral observations to typecast workers as one of *Spammer*; *Sloppy*; *Incompetent*; *Competent*; or *Diligent*. Here the authors take a keen interest in designing this typology with an aim to attract workers with desirable features, rather than to understand the competencies of the worker population.

As discussed by Dukat and Caton [8], these existing approaches are seldom applied to ascertain actual worker competencies. They merely serve as an indicator for whether a worker is likely to possess the required ability to complete a microtask successfully, and whether a worker is trustworthy. In this paper, we present an understanding of the diversity in competence of individual crowd workers.

In closely related works by Kosinski and Bachrach et al., the authors measured the performance of crowd workers on a standard IQ questionnaire [1, 24]. The authors however, discuss factors that effect the overall performance such as composition of the crowd, reputation of workers and monetary rewards. Finally, the authors discuss an approach to aggregate responses from crowd workers to boost performance. While in these works the authors show that aggregating responses from crowd workers is a profitable approach, in this paper, we are more interested in the individual competence of workers, and therefore adopt a more granular view of responses.

Previous works have highlighted the importance of building tools that support crowd work from the perspective of workers, in order to address the power asymmetry in existing crowdsourcing platforms such as AMT [17, 31, 32]. In addition to this, Kittur et al. identified *facilitation of learning* as an important next step towards building a bright future for crowd work [23]. Complementary to these initiatives, we propose the use of self-assessments in pre-selection of workers to aid requesters in recruiting the desired crowd. In the future, we can explore the potential use of self-assessments to help workers increase their self-awareness, identify and potentially facilitate learning where their skills are lacking. Thus, we believe that there can be promising new directions based on leveraging workers' self-assessments to support and improve crowd work in various domains.

11 CONCLUSIONS AND FUTURE WORK

Our work presented in this paper has important implications on paid microtask crowdsourcing systems, since we show that there is a disparity in the crowd regarding the metacognitive ability of workers. This hinders the performance of workers and deprives learning. Through our experiments and results presented in this work, we see evidence of the Dunning-Kruger effect in the paid crowdsourcing paradigm. By studying the impact of inherent task difficulty in the logical reasoning task, and exploring three hypotheses, we reached the following main conclusions and novel contributions.

(i) The important contribution that our work adds to existing literature on self-assessment is the impact of task difficulty on the Dunning-Kruger effect among crowd workers. In tasks with relatively lower difficulty (lower grades), we clearly observe the Dunning-Kruger effect. However, we note that with an increase in grade levels, competent workers also tend to gradually shift towards over-estimation of their ability and performance. This is explained by the fact that the higher grades go beyond the capabilities of even the competent crowd workers (research questions **RQ#1, RQ#2**).

(ii) The capability of a worker to accurately self-evaluate is an integral aspect of the worker's competence. Through our rigorous evaluation in tagging, sentiment analysis and image validation tasks, we have observed that crowdsourcing microtask requesters can benefit by operationalizing workers' self-assessments as a means of assessing their competence rather than relying solely on their performance in worker pre-selection phases (**RQ#3**). We find that workers pre-selected using our proposed approach exhibit a significantly higher accuracy, than those that are obtained using a traditional pre-screening method.

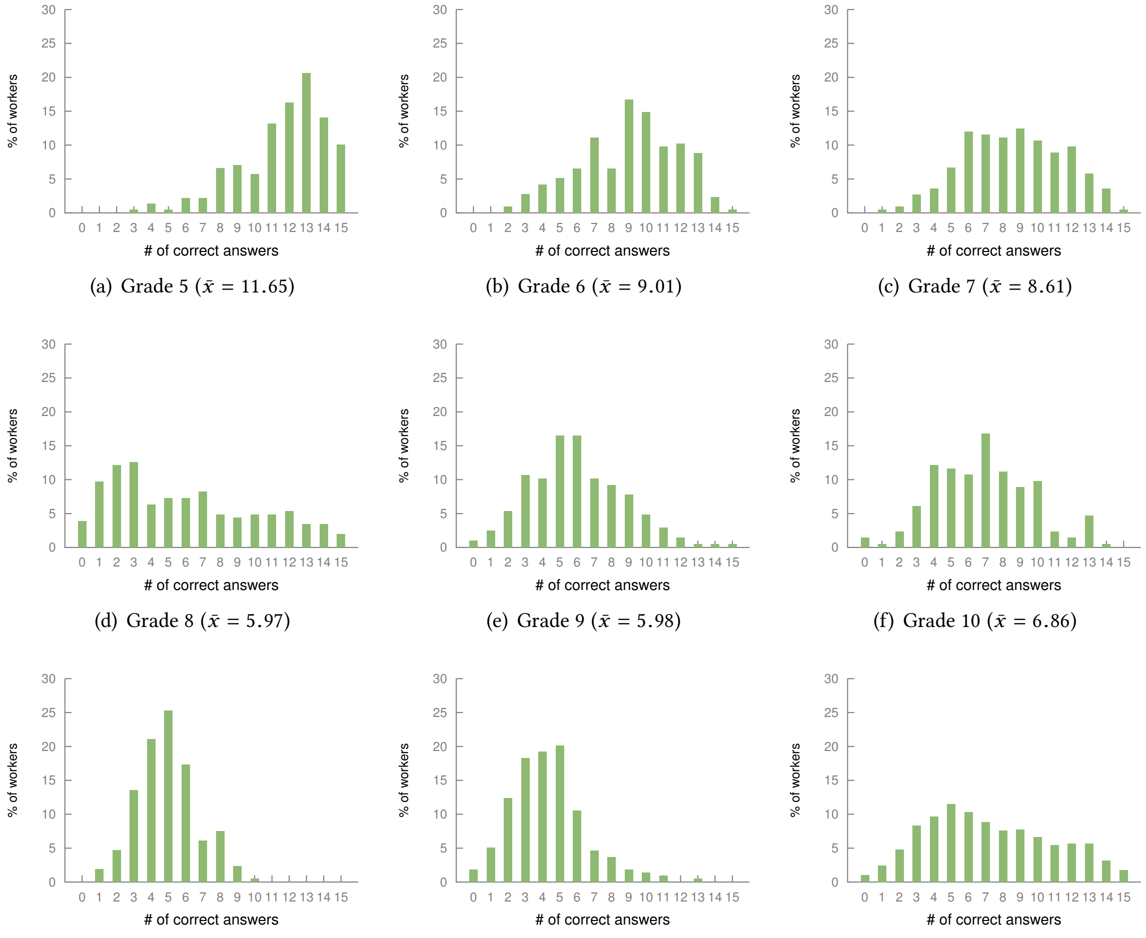
Our findings enrich the current understanding of crowd work and structuring workflow. In the imminent future, we will investigate the use of self-assessments to help workers increase their self-awareness, identify and potentially facilitate learning. We will also test the applicability of our proposed approach across different domains of crowd work. We will investigate the resilience of self-assessment based pre-screening to adversarial attacks, in comparison to traditional pre-screening methods.

A APPENDIX : ANALYSIS OF CONFOUNDING VARIABLES

In this appendix, we provide a thorough analysis of workers that participated in the graded tasks (G5-G12) in Study I and the confounding variables.

A.1 Other Demographics and Worker Performance

The overall *performance* of a worker in each grade is measured as the number of correctly answered questions. We investigated the transition of the workers performance from low to high grades; in the *5th grade*, the majority of workers achieved a score of 13 correct answers out of 15. This declined to 5 in case of the *12th grade*, as shown in Figure 13(h). Figure 13(i) depicts the overall performance of workers across all grades.



(a) Grade 5 ($\bar{x}$ = 11.65)

(b) Grade 6 ($\bar{x}$ = 9.01)

(c) Grade 7 ($\bar{x}$ = 8.61)

(d) Grade 8 ($\bar{x}$ = 5.97)

(e) Grade 9 ($\bar{x}$ = 5.98)

(f) Grade 10 ($\bar{x}$ = 6.86)

(g) Grade 11 ($\bar{x}$ = 4.93)

(h) Grade 12 ($\bar{x}$ = 4.31)

(i) Overall (μ = 7.21)

Fig. 13. The distribution of workers based on the number of correct answers (with a maximum of 15). '$\bar{x}$' represents the mean performance of workers in each corresponding grade, and 'μ' represents the mean performance of workers across all grades.

We found deteriorating competence of the worker pool with the progressive difficulty of the grade levels. This difference in worker performance with respect to the grades is significantly different between all the grades (*p<.01*), with the exception of (*G6*, *G7*), (*G8*, *G9*), and (*G8*, *G10*) using multiple t-tests and Bonferroni correction for type-I error inflation.

If we assume that in order to pass the test, crowd workers are required to score over 50% (>7/15; more than 7 out of 15), we note that the percentage of workers that pass is a monotonically decreasing function from G5 (about 93%) through G12 (about 8%). Considering the passing score to be over 50% (>7/15) in each grade, we note that on average the crowd workers pass only G5, G6 and G7. Hence, crowd workers while working individually in this task setting, can be said to be capable of passing the 7^{th} grade on average.

A.1.1 Majority Voting vs. Correct Answers. Considering that *majority voting* is a widely used scheme for aggregating judgments from crowd workers, we analyzed the correlation of cases where the *majority voting* scheme corresponds to the correct answer to a question across the different grades.

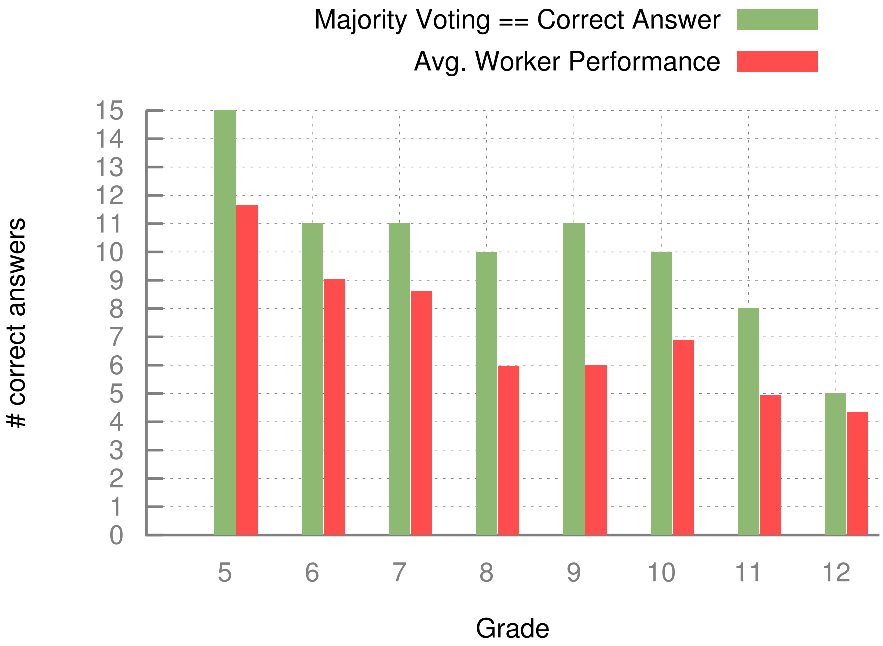


Fig. 14. Number of correct answers based on majority voting and the average performance of workers in each grade.

In Figure 14, we show the number of times that an answer to a question based on the majority voting is actually correct, in contrast to the case where the answer emerging from majority voting happens to be incorrect for the different grades. With the gradual increase in the difficulty level from G5 through G12, such agreement between the majority voted answer and the correct answer decreases rapidly. In the *5th grade* majority voting always corresponds to the correct answer, whereas in the *12th grade* it leads to the correct answer in only 5 cases. Considering that a passing grade for the crowd is 50%, we note that crowd workers collectively, can be said to be capable of passing the 11^{th} grade. Our findings here, align with studies of collective wisdom in the crowd that show an improvement in accuracy over crowd workers working individually [1].

A.1.2 Crowdsourcing Channels. The workforce of CrowdFlower participating in tasks on the platform, originates from various third-party channels[13]. We analyzed the performance of the workers from different channels, and their contribution to the task completion. Our findings are presented in Table 2. We note that nearly 90% of workers in our tasks come from two channels, namely `clixsense` and `neodev`. We found that across the different channel groups, Levene's test for homogeneity of variances was not violated; $F(6, 1663) = 0.7311$, $p = 0.6246$. Thus, we conducted a one-way between workers ANOVA to compare the effect of the top 7 *channel* groups (see Table 2) on the performance of workers (across all grades). We found no significant difference in the performance of workers at the $p < .05$ level for the 7 channel-group conditions [*F(6,1663)=1.1722, p=0.32*].

[13] https://www.crowdflower.com/labor-channels/

Post-hoc comparisons using the Tukey HSD test confirmed the lack of significant difference between any two of the 7 channel-group conditions.

Table 2. Top 7 third-party channels through which workers participated in our tasks from G5-G12. The left-half of the table shows the percentage of workers that account to specific grades. The right-half of the table shows the average number of correctly answered questions.

channel	*% of workers*			*Avg. Performance*		
	5th	*12th*	*Overall*	*5th*	*12th*	*Overall*
clixsense	44.30	47.49	45.12	11.69	3.96	7.16
neodev	37.72	34.70	38.69	11.70	4.45	7.10
elite	9.21	4.57	7.08	11.57	5.90	7.1
tremorgames	3.51	2.74	2.29	12.50	4.17	8.76
getpaid	1.32	1.37	0.68	1.03	3.67	6.7
gifthunterclub	0.88	2.74	0.6	11.00	6.00	5.13
instagc	0.88	1.83	0.23	1.5	4.25	7.39

A.1.3 Worker Attributes: Age and Education. **Age.** We collected responses from the crowd regarding their age group, to investigate the influence of age on the performance of workers. Table 3 presents our findings. The biggest group of workers lies in the age group of 26-35 years, which accounts for nearly 42% of all crowd workers in our task. Due to homogenity of variances, we conducted a one-way between workers ANOVA to compare the effect of *age* on the performance of workers (across all grades) in the 5 age-group conditions presented in Table 3. However, there was no significant effect of *age* on the performance of workers at the *p<.05* level for the 5 age-group conditions [$F(4, 1727) = 0.19, p = .94$].

Table 3. Distribution of workers based on their age group and performance (average number of correctly answered questions) for the different grades G5, G12 and overall (G5-G12).

Age Group	*% of workers*			*Avg. Performance*		
	5th	*12th*	*Overall*	*5th*	*12th*	*Overall*
18-25	27.19	23.29	25.69	11.18	4.45	7.22
26-35	43.42	37.90	41.74	11.52	4.33	7.15
36-45	18.42	21.00	19.98	12.55	4.09	7.36
46-55	7.46	13.70	9.64	11.88	4.50	7.19
> 55	3.51	4.11	2.94	11.88	4.11	7.31

Education. In Table 4 we present the distribution of workers based on their *educational qualifications* and their performance. A majority of workers have a *Bachelor's* degree, while a small percentage has a *Doctoral* degree. We found that Levene's test for homogenity of variances was not violated across the different educational qualification groups; $F(9, 1722) = 1.5274$, $p = 0.1327$. Thus, we conducted a one-way between workers ANOVA to compare the effect of *education* on the performance of workers (across all grades) in the 10 educational qualification conditions presented in Table 4. There was a significant effect of educational qualification on the performance of workers at the *p<.001* level for the 10 conditions [$F(9, 1722) = 4.889, p = .0001$]. Post-hoc comparisons using the Tukey HSD test indicated that the mean performance of workers with the educational qualification of 'Some high school (no diploma)' *(M=5.62, SD=3.58)* and 'Technical/Vocational training' *(M=6.09, SD=3.28)* was significantly different than

Table 4. Distribution of workers based on their educational qualifications and performance (average number of correctly answered questions out of 15).

Education	*% of workers*			*Avg. Performance*		
	5th	*12th*	*Overall*	*5th*	*12th*	*Overall*
No schooling	0.00	0.46	0.17	0.00	4.00	5.67
Some high school (no diploma)	2.63	6.39	3.98	13.33	3.79	5.62
High school	14.91	11.42	11.43	11.06	3.84	6.80
Some college (no degree)	11.84	13.70	12.76	11.56	4.50	6.99
Technical/vocational training	3.95	7.76	5.72	11.44	4.53	6.09
Associate degree	3.95	5.48	4.68	11.11	4.42	6.99
Bachelor's degree	34.65	24.66	34.64	11.42	4.44	7.67
Professional degree	8.77	11.87	8.43	12.20	4.04	6.95
Master's degree	17.98	14.61	16.69	12.17	4.78	7.74
Doctorate	1.32	3.65	1.50	13.67	3.75	7.31

that of workers with the educational qualification of Bachelor's *(M=7.67, SD=3.59)* or Master's degree *(M=7.74, SD=3.69)*. However, there was no significant difference between the workers with other educational qualifications.

A.2 Task Completion Time

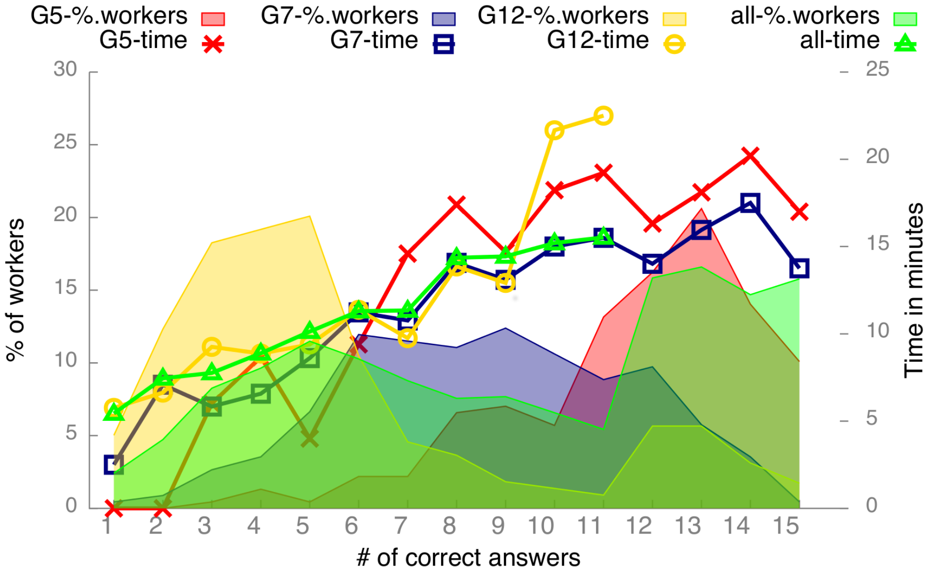


Fig. 15. Distribution of workers based on their performance (number of correct answers on x-axis) and task completion time (in minutes on y2-axis) for the different grades.

Figure 15 presents the correlation between the *task completion time* for a grade and the resulting performance of the workers. It is evident that there is a direct correlation between the performance score of workers and the *completion time*. We found a positive correlation between the performance and completion time up to grade 7. The correlation coefficients (as measured based on Pearson's **r**) for grades 5 upto 7, are 0.68, 0.35, and 0.47 respectively. In case of grades higher than 7, we found a weak negative correlation. Based on these observations, we reason that if a worker is not capable of solving a task, the task completion time does not influence the corresponding performance in the given task.

An interesting insight from Figure 15 is that poorly performing workers depict a significantly lesser *task completion time* than the best performing workers. While this difference varies between grades, the overall difference between workers with an average number of correct answers below 5, and those with more than 10 correct answers, is over 10 minutes.

ACKNOWLEDGMENTS

This research has been supported in part by the European Commission within the H2020-ICT-2015 Programme (AFEL project, Grant Agreement No. 687916).

REFERENCES

[1] Yoram Bachrach, Thore Graepel, Gjergji Kasneci, Michal Kosinski, and Jurgen Van Gael. 2012. Crowd IQ: aggregating opinions to boost performance. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 535–542.

[2] David Boud. 2013. *Enhancing learning through self-assessment*. Routledge.

[3] Katherine A Burson, Richard P Larrick, and Joshua Klayman. 2006. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of personality and social psychology* 90, 1 (2006), 60.

[4] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. 2016. Chain Reactions: The Impact of Order on Microtask Chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 3143–3154.

[5] Lydia B Chilton, John J Horton, Robert C Miller, and Shiri Azenkot. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 1–9.

[6] D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing – The Case of Amazon MTurk. In *24th International Conf. on World Wide Web (WWW)*. ACM, 238–247.

[7] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.

[8] Christoph Dukat and Simon Caton. 2013. Towards the Competence of Crowdsourcees: Literature-based Considerations on the Problem of Assessing Crowdsourcees' Qualities. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, 536–540.

[9] David Dunning. 2011. The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. *Advances in experimental social psychology* 44 (2011), 247.

[10] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.

[11] Joyce Ehrlinger and David Dunning. 2003. How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of personality and social psychology* 84, 1 (2003), 5.

[12] Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes* 105, 1 (2008), 98–121.

[13] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*. ACM, 105–114.

[14] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training Workers for Improving Performance in Crowdsourcing Microtasks. In *Proceedings of the 10th European Conference on Technology Enhanced Learning. EC-TEL 2015*. Springer, 100–114.

[15] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 218–223.

[16] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*. 1631–1640.

[17] Lilly C Irani and M Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620.

[18] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk. In *A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, Michigan, USA, August 4-8 2011*. Association for Information Systems.

[19] Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval*. Springer, 165–176.

[20] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 1941–1944.

[21] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2583–2586.

[22] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.

[23] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[24] Michal Kosinski, Yoram Bachrach, Gjergji Kasneci, Jurgen Van-Gael, and Thore Graepel. 2012. Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 151–160.

[25] Marian Krajc and Andreas Ortmann. 2008. Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology* 29, 5 (2008), 724–738.

[26] Joachim Krueger and Ross A Mueller. 2002. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of personality and social psychology* 82, 2 (2002), 180.

[27] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[28] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2015. Peer and self assessment in massive online classes. In *Design Thinking Research*. Springer, 131–168.

[29] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 21–26.

[30] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 234–243.

[31] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 224–235.

[32] David Martin, Jacki OfiNeill, Neha Gupta, and Benjamin V Hanrahan. 2016. Turking in a Global Labour Market. *Computer Supported Cooperative Work (CSCW)* 25, 1 (2016), 39–77.

[33] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazonfis Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[34] Tyler M Miller and Lisa Geraci. 2011. Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning* 6, 3 (2011), 303–314.

[35] Edward Newell and Derek Ruths. 2016. How One Microtask Affects Another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. ACM, 3155–3166.

[36] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation* 11, 11 (2011).

[37] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[38] Thomas Schlösser, David Dunning, Kerri L Johnson, and Justin Kruger. 2013. How unaware are the unskilled? Empirical tests of the fisignal extractionfi counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *Journal of Economic Psychology* 39 (2013), 85–100.

[39] Barry Schwartz. 2004. The paradox of choice: Why less is more. *New York: Ecco* (2004).

[40] Barry Schwartz and Andrew Ward. 2004. Doing better but feeling worse: The paradox of choice. *Positive psychology in practice* (2004), 86–104.

[41] Han Yu, Zhiqi Shen, Chunyan Miao, and Bo An. 2012. Challenges and Opportunities for Trust Management in Crowdsourcing. In *2012 IEEE/WIC/ACM International Conferences on Intelligent Agent Technology, IAT 2012, Macau, China, December 4-7, 2012*. IEEE Computer Society, 486–493.