IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2022, accepted 2 September 2022, date of publication 16 September 2022, date of current version 26 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3207292

RESEARCH ARTICLE

What Quality Control Mechanisms do We Need for High-Quality Crowd Work?

MARGERET HALL[1], MOHAMMAD FARHAD AFZALI[2], MARKUS KRAUSE[3], AND SIMON CATON[4]

[1]Department of Strategy and Innovation, Wirtschaftsuniversität Wien, A-1020 Vienna, Austria
[2]College of Information Science and Technology, University of Nebraska Omaha, Omaha, NE 68182, USA
[3]Brainworks.ai, Berkeley, CA 94501, USA
[4]School of Computer Science, University College Dublin, Dublin, D04 V1W8 Ireland

Corresponding author: Simon Caton (simon.caton@ucd.ie)

ABSTRACT Crowd sourcing and human computation has slowly become a mainstay for many application areas that seek to leverage the crowd in the development of high quality datasets, annotations, and problem solving beyond the reach of current AI solutions. One of the major challenges to the domain is ensuring high-quality and diligent work. In response, the literature has seen a large number of quality control mechanisms each voicing (sometimes domain-specific) benefits and advantages when deployed in largescale human computation projects. This creates a complex design space for practitioners: it is not always clear which mechanism(s) to use for maximal quality control. In this article, we argue that this decision is perhaps overinflated and that provided there is "some kind" of quality control that this obviously known to crowd workers this is sufficient for "high-quality" solutions. To evidence this, and provide a basis for discussion, we undertake two experiments where we explore the relationship between task design, task complexity, quality control and solution quality. We do this with tasks from natural language processing, and image recognition of varying complexity. We illustrate that minimal quality control is enough to repel constantly underperforming contributors and that this is constant across tasks of varying complexity and formats. Our key takeaway: quality control is necessary, but seemingly not how it is implemented.

INDEX TERMS Quality control, human computation, natural language processing, image recognition, crowd work.

I. INTRODUCTION

Human computation such as crowd labour powers the ability to access, exploit, and disseminate knowledge at scale. As a domain and artificial computational paradigm, it has received significant attention from researchers in exploring the "best" manners to leverage and efficiently utilise large quantities of online workers. However, there are still aspects of crowd work where many end-users, i.e. employers of crowd workers, are still in need of structured guidance and best practice recommendations. In this article, we focus on just one of these key decisions: choosing the "right" mechanism(s) for quality control, which is often a key consideration for any large(r) project involving human computation [1], [2]. This is because data quality is the cornerstone assumption of information and computer technology [3], [4] and ultimately, crowd workers generate data for use in other systems and contexts. We argue that it is unclear how to select from the many quality control mechanisms discussed throughout the literature and seek to provide guidance for practitioners utilising crowd platforms.

Questions around digital work, and specifically quality in digital work, are not without founding. The 2019-2020 outbreak of the Coronavirus (COVID-19) pandemic demonstrated the universality of working remotely and from home. More to the point, widespread unemployment has millions scrambling at the margins of the workforce; the value of crowdsourcing is proven in such circumstances as crowd labour has turned from an option to the only option for certain individuals [3]. The lack of working in a physical setting opens up a huge and ready workforce and indeed

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

crowdsourcing has been leveraged for research and industry, such as its use for medical purposes – i.e., [4] – during the COVID-19 pandemic.

The discussion around quality control often emphasises several different challenges, and different approaches to quality control in crowd work address different combinations of these challenges. To portray the complexity in selecting appropriate quality control mechanisms, we first need to discuss the rationale behind integrating quality control mechanisms into crowd work, i.e., the challenges in crowd work they seek to address, and some of the high-level implications that accompany different mechanisms.

The challenge of "underperforming" workers is primary. This is what quality control mechanisms are seeking to hinder, discourage, and often punish. In this article, we intentionally do not refer to these workers as "spammers" or "cheats" as there is often no means by which to categorise their intent [5], i.e. we cannot distinguish between intentional low quality work, and a lack of training for the task at hand. We discuss common mechanisms for quality control in Section-II; broadly speaking, skill tests [6]; pre-set qualifications and benchmarking against metrics like historical solution acceptance [7], [8]; trust-based models [9], [10], [11]; disguised gold standard questions [12]; and machine learning-driven trackers [13] have all been proposed by researchers for quality control. Our objective is not a systematic review of all mechanisms, nor playing one off against another, but rather contextualising their impact more holistically towards understanding how well, when, and why specific quality control mechanisms work [14]. This becomes particularly important when contextualised with findings like [15] who, against expectations, found that worker quality is stable over time.

Secondly, we must consider the implications that any quality control mechanism has. We can categorise these implications into the broader areas of 1) respect, 2) relationship, and 3) ethics. Yet, we also recognise that this is not an exhaustive categorisation. These categories are important because researchers are now recognizing that implemented quality control mechanisms must not only be effective but also should be respectful of the worker [16], [17], [18]. As such, it is not just a case of identifying which quality control mechanism(s) will maximise specific notions of quality. While the choice of mechanism(s) clearly impacts the task design, it also affects the relationship between worker and requester, especially if the worker perceives the quality control mechanism(s) in use to be (overly) strict or unfair. [19] have also noted that some quality control methods may pose ethical dilemmas. Thus, for practitioners, there is not only the question of which quality control mechanism is "right" but also the consideration of any (in)direct implications.

Related to this is the proposition that crowd labour remains in a "perpetual beta" state [20]. The wealth of approaches in the literature makes it clear that much research has been conducted to support both workers (in validating their work, ensuring they are qualified to do it, etc.) and employers (seeking to ensure high-quality solutions). This article therefore recognizes that there is a wealth of choice for researchers and requesters [21], but that this choice adds complexity to the design and implementation of crowd work. We note that several researchers have highlighted some shortcomings in the literature (see *Related Work*) that we seek to either address or provide more context experimentally. We do so by proposing the following research questions (RQ):

RQ1: what characterisations of response quality can be linked to different quality control mechanisms?

RQ2: what impacts of task design have larger effects on response quality then different mechanisms of quality control?

In RQ1, the working hypothesis is that the format of the quality control mechanism is critical for achieving specific notions of quality. Building on this, RQ2 attempts to disentangle questions surrounding when to inject quality control mechanisms in the task. We seek to experimentally address the impact of differing quality control methods on contributors' response quality (RQ1). Also, reflecting on [22] we explore the role of task and interface design in the quality and accuracy of the tasks (RQ2).

Aligned to the two research questions, we present two factorial design experimental studies (see *Study Design*). The first, a 3 × 5 design explores 3 different task complexities in language processing using 5 quality control methods. It emphasises the impact of quality control mechanisms vs. task complexity when considering response quality. The second, a 3 × 2 × 2 design explores 3 quality control treatments with 2 information highlighting approaches and 2 task ordering effects within a simple image recognition task. It emphasises how aspects of task design impact response quality. In undertaking these experiments, we make the following observations (see *Results*).

1) Consistently underperforming workers were repelled by the simple announcement of a quality control mechanism, regardless of what that mechanism was, or in fact if one was actually present or not.
2) There was no statistically significant difference between the quality control mechanisms applied.
3) Subtle considerations in the task design (e.g. making key text bold, and the order tasks are performed) are more impactful on quality than the effect associated with a quality control mechanism.

Considering these observations, we argue that the wealth of choice for researchers and requesters in achieving quality control [21] only adds complexity to the design and implementation of crowd work. Instead, we provide a set of recommendations for practitioners based on the following contributions (see *Discussion and Conclusions*):

1. Quality control mechanisms: we highlight that in some cases, the presence of a quality control measure alone is sufficient to ensure high(er) quality solutions. This is key for crowd requesters, as Rzeszotarski and Kittur [21] note requesters must make difficult trade-offs depending on the quality control method they use, yet our results illustrate this may be an over-emphasized issue: we could not observe discernible differences between increasingly more sophisticated measures (RQ1 and experiment 1). Similarly, as Difallah *et al.* [23] suggest, discouraging low quality work (or ''cheaters'') is better than controlling the quality of results.

2. *Task design*: via RQ2 and experiment 2, we provide insights into task design aspects and their relationship with observable differences in quality. Newell and Ruths [24] state that intertask effects could create a systematic bias (if left unchecked), and they note the importance of task design. Experiment 2 corroborates this finding, illustrating the impact of subtle differences in task framing and feedback. Similarly, Cai *et al.* [25] note the importance of task ordering aligned to cognitive load. Experiment 2 explores this aspect of task design. Finally, experiment 2 also corroborates the findings of Chandler and Kapelner [26] who note the relationship between task framing and solution quality. Yet, experiment 2 further extends these findings by exploring a rich design space: clarity of instructions vs. cognitive ordering vs. quality control-based feedback mechanisms.

II. RELATED WORK

Quality control for crowd platforms is a highly-studied phenomenon as quality is a major attribute of the crowd [1], [27], [28]. Applications outside of industry abound including creative endeavours [29], policy and budget deliberations [30], [31], [32], open collaboration platforms [33], and the academic research community [34]. Literature suggests there are several factors that may work quality including worker demographics [35] and personality traits [36]. (Under)performance can also be linked to the requestor; recent works finds a high percentage of workers complain about the task instructions being unclear and the language in the task description being difficult to understand. Workers see task clarity as playing a major role in their performance [37]. Task complexity, while being subjective, can be measured by visual appearance and language used in task description [38].

Various measures for assuring quality and authenticity have been proposed. Corrigan-Gibbs *et al.* [39] conducted two experiments with student participants and with MTurk respectively, testing the difference between honour codes and a serious warning message. They found a 50% decrease in cheating in both student and MTurk environments when replacing a traditional honour code with a strict warning. The authors suggest that informing participants of the negative consequences of an action in a warning results in a lesser tendency of doing it. In a pair of experiments on MTurk, Kittur *et al.* [40] first asked MTurkers to rate Wikipedia articles regarding their accuracy, writing quality, neutrality, structure, and the overall quality of the article. To verify that workers had read the article, they filled in a text box with suggesting improvements to the article. Based on the five metrics the authors found no correlation between the MTurkers ratings and the actual Wikipedia administrators. In a subsequent experiment the authors made slight modifications and added both subjective and objective questions. Users were first asked verifiable, quantitative questions and then to rate the article. They also provided 4-6 keywords as a summary for the article. The results of the subsequent experiment demonstrated a significant positive correlation between the workers' ratings and the Wikipedia admin ratings. The combined findings indicate the utility of combining objective and subjective tasking in micro-task markets [40].

A recognised attribute of crowdsourcing platforms is that the platforms do not identify workers nor guarantee the quality of the work, which can contribute to in the unreliability of the system [23], [41]. In their work, Difallah *et al.* categorized 'cheaters' *a priori* and *posteriori* and discuss anti-adversarial techniques for encountering them. They suggest that sophisticated task formulation as a suitable obstacle for cheaters. Requesters' main goal is receiving high-quality, done work thus discouraging 'cheaters' from doing a task in the first place is more goals compatible than controlling the quality of completed tasks. However, more sophisticated or complicated task structuring increases the burden on the requester. They propose traditional anti-spamming techniques such as CAPTCHA as sufficient barriers to 'cheaters'. Several common approaches for quality control exist that are discussed next.

A. PRE-SELECTION METHODS

Pre-selection mechanisms have two main branches which are differentiated as ''up-front task design'' and ''post-hoc result analysis'' [42] to control work quality in a crowdsourcing context. Researchers have utilized various techniques to apply pre-selection methods. Crowdsourcing platforms generally provide a mechanism for requesters to pre-select contributors based upon specific task requirements or requester preferences. Geiger *et al.* [43] typify pre-selection as ''a means of ensuring a minimum ex-ante quality level of contributions.'' Otherwise stated, a requester uses a pre-selection process like a test as a risk mitigation technique against poor-quality solutions. Namely requesters screen potential contributors based upon the demonstration of certain knowledge, skills, or attributes via platform-specific process.

Pre-selection is typically performed via multiple-choice tests, which Oleson *et al.* [12] examined and subsequently criticized due to a faulty key assumption: if the contributor passes the test, they will then perform the task well even in the absence of direct or tangible incentives to do so. Likewise, contributors who fail the test may be banned from the task though not necessarily for the right reasons. Gadiraju *et al.* [44] found that identifying workers' behavioural traces can help with classifying worker in different types that will then improve the quality of work produced significantly. This improvement was more significant in high complexity tasks.

Self-assessments as a pre-selection technique produced promising results in providing a strong indicator for workers' competence and potential performance [45]. This method is simple to implement and has been found to perform well. Because of additional unremunerated efforts required and the demonstration in advance of credentials in this design, preselection via qualification tests also likely acts as a barrier to "spammers" [45]. This is however a double-edged sword as diligent contributors also may not select the task due to increased unremunerated effort or missing credential on their part. Answers to qualification tests or generic credentials may also be shared amongst users, which reduces effectiveness of the QA method [46], [47].

B. QUALIFICATION TESTS

Qualification tests can be used to not only determine the abilities of a contributor but also reveal workers' basic attributes, as this information is often not available in advance [28], [48]. Depending on the requirement, qualifications can also capture demographic properties of a contributor, for example, their geographical location. This does, however, massively distort the concept of 'qualification' if personal attributes are considered.

Like the notion of qualification tests, initial screening questions based on task attentiveness can be employed to minimize 'click-through' behaviours [49]. Such measures aim to ensure that contributors dedicate attention to key elements of information, like reading and understanding the task's instructions.

C. IN TASK QUALITY CONTROL

Ipeirotis *et al.* [10] and Sheng *et al.* [50] proposed inferring a level of trust in the contributor linked to the accuracy of their solutions. Trust, however, quickly becomes a complex and nuanced topic highly specific to the context in which it is considered. Also, as an inherently intangible and intransitive construct trust is challenging to quantitatively establish even though a measurement of trust is a key aspect for (automatically) approximating a contributor's propensity for reliable or diligent work. Thus, Kern *et al.* [51] propose proxying trustworthiness based on prior experience. It is reported that worker-requester trust has a positive impact on the reliability of the crowd work [19]. The authors suggest that one way of enhancing worker-requester trust is to flag and scrutinize workers with sub-optimal responses rather than rejecting their work and not paying them.

To provide a basis for comparing and estimating contributor reliability Kern *et al.* [51] redundantly scheduled tasks for multiple contributors. While this method demonstrated yielding high-quality solutions without careful management the method quickly becomes expensive in terms of the direct costs of redundantly issuing tasks and indirect costs of additional effort needed to assess solution quality. Similarly, with respect to "rejected" answers, such methods can have other adverse effects with respect to contributors who have acted diligently.

Gold standard questions are frequently used on MTurk to assess solution quality and contributor attributes. In their approach, Oleson *et al.* [12] inject known solutions into the task as subtasks and contributors receive instant feedback on the accuracy of their performance. The presence and quality of these subtasks enables the accuracy of a given contributor to be estimated in-task. As it is in-task, it also helps to improve the quality of workers' solutions by providing an explanation of why the solution is incorrect. The approach, however, is inappropriate for tasks that rely on forms of subjectivity as the design requires a finite set of definite answers. However, such a mechanism also provides a basis to train contributors, enabling self-evaluation of performance through feedback. The latter facilitates an integral element in the definition of competence: the evaluation of self-efficacy.

D. THE ROLE OF MOTIVATION

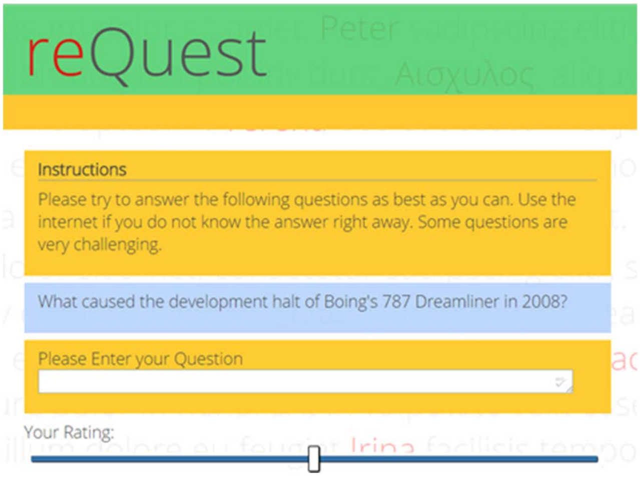
Completing meaningful tasks leads to motivation in the workplace [52]. Meaningful in this context implies that the worker is both doing work with purpose and receiving acknowledgements for accomplishments [53]. Chandler and Kapelner [26] transferred these findings into the crowd environment showing an interdependency between how a task is framed and outcome in terms of work output. Motivating task rationale in terms of expressing a purpose and higher goal led to a significantly higher willingness for participation and quantity of output.

Quality control is a dimension of Quinn and Bedersen's human computation classification [54]. They caution that even motivated users might cheat or sabotage the system. We argue that the rationale behind subpar performance is that the motivation typically studies is extrinsic rather than intrinsic motivation [55]. Ke *et al.* [56] investigated the role of intrinsic motivation in adoption of Enterprise Systems among employees from the lens of self-determination theory. The authors investigated if inducing intrinsic motivation results in better and smoother adoption of Enterprise Systems in an organization. Their findings suggest that individuals' intrinsic motivation should be enhanced to adopt or explore new systems.

Ryan and Deci [57] define extrinsic motivation as "the performance of an activity in order to attain some separable outcome" or the performance of an activity to avoid punishment. Zhao *et al.* [58] studied the role of extrinsic motivation in having individuals share their knowledge in Q&A sites. They argue that while extrinsic motivation, when used as a reward, could help increase participation and knowledge sharing it might also interact with intrinsic motivation, impacting self-esteem and self-actualization. It is unknown to which degree this interaction between extrinsic motivation and intrinsic motivation impacts quality control, which weighs towards the punishment end of extrinsic motivation, in the crowd.

E. SUMMARY OF RESEARCH GAP

There is a significant amount of work on both assessing and trying to ensure the quality of crowd work. These approaches typically reside prior to the undertaking of a task (e.g. qualification tests), or in-task (e.g. gold standards, redundant task scheduling). Choosing the "right" measure for a given task, however, is challenging, as many researchers have proposed many different quality control / assurance measures [21].



reQuest

Instructions

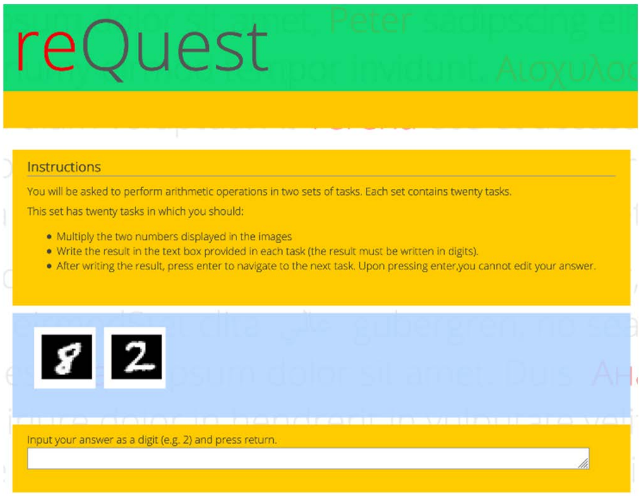
Please try to answer the following questions as best as you can. Use the internet if you do not know the answer right away. Some questions are very challenging.

What caused the development halt of Boing's 787 Dreamliner in 2008?

Please Enter your Question

Your Rating:

FIGURE 1. Crowdsourcing interface demonstrating the question answering task. The basic interface is identical for all tasks. The rating slider (bottom) is only visible for our Raters when they judge the quality of a response.



reQuest

Instructions

You will be asked to perform arithmetic operations in two sets of tasks. Each set contains twenty tasks.

This set has twenty tasks in which you should:

- Multiply the two numbers displayed in the images
- Write the result in the text box provided in each task (the result must be written in digits).
- After writing the result, press enter to navigate to the next task. Upon pressing enter,you cannot edit your answer.

Input your answer as a digit (e.g. 2) and press return.

FIGURE 2. Crowdsourcing interface for the second experiment, here illustrating that the worker should recognize and then multiply the two images. Shown is a non-bold verb, control group view.

This creates a complex task design space for requesters and researchers wishing to leverage crowd platforms. Ultimately, they have to find a balance between reducing risks, and building trust [59]. However, there is not yet a coherent or systematic review of quality control measures in the literature to help guide these design choices. In this paper, we shed some light on the roles that the exact quality control measure has and link this to specific aspects of task, interface, and instruction design.

III. STUDY DESIGN

This work leverages two studies. The first tests the impact of different quality control measures upon worker response quality. The second evaluates if there is an effect concerning when quality control is in place. We concentrate specifically on underperformance, rather than increasing the performance of already-acceptably performing workers. The first experiment considers the domain of natural language processing. The second experiment considers Roman numeral image recognition and numeracy (basic arithmetic). The main interface for contributors is identical across both experiments and all tasks. Figures 1 and 2 show a screenshot of the user interface for two conditions of the experiments.

A. PARTICIPANTS

We recruited all contributors for the first experiment via Crowd Flower. The second experiment was conducted with its successor platform – Figure Eight. The reasoning behind this is due to a rebranding of Crowd Flower to Figure Eight in between the two experiments taking place.

Contributors were prompted to access the task on our own web page. This allows for confounding variable control, personalised feedback, and performing our own quality control. The website created a unique code that contributors use to receive their payment through the Crowd Flower interface after completing the task. The user interface (Figure 1 and Figure 2) was identical for all conditions across both experiments. The same interface was used for collecting human judges' quality ratings in experiment 1.

We used between-group designs where each task had its own population (groups had no overlap among populations). To ensure this, we used IP-tracking and browser fingerprinting to ensure that contributors do not contribute to more than one condition, as well as corresponding constraints specified via the Crowd Flower and Figure Eight platforms.

In Figure Eight, 60% of the workers are male, and most of the workers are between the ages 18 and 34 years, aligning with recent assessments of crowd labour participants [58]. For this study, we did not collect demographic information as it did not serve the aim of the experiments. Only hashes of IP's and browser fingerprints were stored to maintain participant privacy. Table 1 and Table 2 display contributor distribution of the experiments in this study.

B. EXPERIMENT 1

1) DIFFERENTIATED QUALITY CONTROL MECHANISMS

Our first experiment investigates three tasks of varying complexity using a three (task complexities) by five (quality control methods) factorial, between-group design. Following the experimental design of [1], the effort for completing each task is as high or higher than for cheating, which should disincentivize constant underperformance. We hypothesize the levels of complexity to be *semantic similarity* (least complex); *question answering* (more complex); and *text translation* (most complex).

Each task is repeated five times with one of five different quality control methods: *none, fake, intro, auto, and wizard*. First, in level (none) we performed no quality control. We announced very prominently in the task description that we use introductory quizzes to check the contributors' qualifications, yet contributors did not actually undertake a test for the (fake) level. The third level (intro) announced an introductory quiz and required contributors to complete the quiz with 80% accuracy which is akin to many qualification tests (*cf.* 'Qualification Tests').

In the fourth level (auto) we added a basic machine learning (ML) system to estimate the quality of response. The ML system uses a three-level scale: *good, acceptable, unacceptable*. This estimate was reported to contributors making it akin to in-task quality control measures. Runge *et al.* have shown that in some natural language tasks response quality can be high accurately estimated by combining the time needed to complete a single request and the number of characters typed [60]. Although the values of both variables and their meaning differ from task to task, an ML classifier can learn the relationship between the two features and subsequently the response quality with minimal training data.

In the auto level, responses were classified into *good, acceptable,* and *unacceptable* levels using a random forest classifier [61]. A random forest classifier was chosen, as tree-based classifiers are less sensitive to outliers or unbalanced sample sets [62]. Classifiers such as support vector machines are more sensitive to such outliers. In the given tasks, outliers are likely i.e., a contributor may open the task and then leave their workstation. Our classifier generated ten random trees using Gini impurity [63] as the split criterion using the python sklearn package [64]. To create the necessary labelled training data for a supervised classifier we classified 90 responses of each task by hand. Responses were selected randomly from the full set and were classified into the three classes until there were 30 samples per class. We stratified the training data randomly, selecting exactly 30 samples per class.

Finally, in the fifth level (wizard) we replaced the ML system with a human observer who determines the response quality using a scale identical to the one used by the ML system. Our objective with this measure is to replicate an expert panel that reviews each solution.

When the classifier or human judge estimates the response quality to be unacceptable a general warning appears that the response might need revision. If the response was acceptable, we did not show a message. For responses assessed as 'good', a message stating that the response was of good quality is shown. These messages appeared as a red text immediately after a contributor responded to a request.

2) MEASUREMENTS

We consider two independent variables (quality control method and task complexity) and one dependent variable: perceived response quality. Two human judges with no overlap across judges between tasks rated each response on a scale from 0.0 (low quality) to 1.0 (high quality) in ten increments to measure perceived response quality. We ensured that the process was blind. Judges were recruited offline, and were not involved in training data generation for the automated feedback nor did they participate in the wizard conditions.

All judges were not informed about the details of the experiment but had experience in crowdsourcing. Judges saw the initial request, answer, and additionally had a slider to rate the response quality (Figure 1) which was not shown on the contributor interface. Responses from all conditions were randomly selected and judges were not informed of which condition a response came from. They were asked to judge performance based on the description of the task as shown to the contributors.

We measure and report the agreement between judges using Krippendorff's Alpha [68]. [66] and [67] illustrate that in a scenario of ten equi-distributed classes with a target Alpha value of 0.8 or higher, a sample size of 293 is sufficient to judge this Alpha level with a p-value < 0.05. As we collected more than 1000 samples, our expected p-value is < 0.005 for an Alpha level of 0.8, which according to Krippendorff is substantial. As illustrated below the provided description was adequate as the observed agreement between judges was substantial with a p-value < 0.05.

We calculated the average perceived response quality for each contributor as our quality measurement. We consider contributors with an average perceived response quality of 40% unacceptable responses, or below 0.6, as constantly underperforming. The value of 0.6 was chosen regarding the ability to recover high quality answers from noise input. A commonly used method for recovering high quality responses from noise human input data is Expectation Maximization. As [65] showed with five raters with an average consistent performance of 0.6 or above a final Cohen's Kappa of 0.9 can be achieved.

Additionally, we measure the correlation between our ML-systems prediction and our human judges. As our data violates the assumptions of the Pearson Product-Moment correlation we use Spearman's ρ. Ground truth data was acquired from the human judgement data. We selected only the samples on which judges achieved full agreement on and selected 30 samples per class. Classifier showed a Cohen's Kappa of > 0.75 in unbalanced test sets resulting in accuracy levels of 0.8 – 0.92 for class balanced test sets. These results are consistent with [61].

In line with [37] and [38], instruction clarity and contributor satisfaction were tested using the built-in metrics provided by Crowd Flower for all three tasks. Upon completion of a task, contributors could opt into a satisfaction survey. Contributors score the task on a 0-5 scale for overall satisfaction, instruction clarity, fairness of test questions, payment, and ease of job. Results of these surveys are reported with each task.

C. EXPERIMENT 2

1) TIME DIFFERENTIATED QUALITY CONTROL

Our second experiment had a one (task difficulty) by three (quality control treatments) by two (information highlighting), by two (ordering), between-group design. The experiment investigated one task of 2 × 20 arithmetic calculations based upon an image recognition exercise that used the MNIST [70] handwritten character dataset. Workers were requested to either add or multiply the values contained within two images (Figure 2 shows the control group). When quality control mechanisms were in place, an additional line to the instructions explicitly announced this.

To select the images from the 42,000 images available within the MNIST dataset, we trained a support vector machine (SVM) offline and prior to the experiment to classify the images. Each image was 28 × 28 grayscale pixels. To train the SVM a 25% (10,500 images) stratified sample was taken and a principal components analysis was undertaken. The first 30 principal components (representing 83.97% variance) were selected and used to transform the remaining 75% test sample of 31,500 images. The SVM uses a Radial Basis Function kernel (with hyperparameters cost=5, epsilon=0.01), and achieves an accuracy of 97.1% on the test set. Although approaches such as Neural Networks can achieve higher accuracies, our objective here was not to build an optimal machine learning model but select images which are "easy" to recognize, i.e. ones where the ML model is very confident in its classification and correspondingly should also be easy for crowd workers. As such, we ranked the classification probabilities for each image in the test set, selecting the ten highest for each digit. The mean probabilities for each set of ten images are as follows: 0: 0.99998; 1: 0.99979; 2: 1.00000; 3: 1.00000; 4: 0.99998; 5: 1.00000; 6: 1.00000; 7: 1.00000; 8: 1.00000; 9: 0.99992. In selecting images in this manner, we sought to reduce the likelihood of ambiguous images in the experiment. Each image was randomly paired with another, giving 50 pairs of images, and we derived the result of addition and multiplication for each pair to facilitate real-time quality assessment and feedback.

Unlike the first experiment, the effort for completing the task is higher than providing some arbitrary response and thus incentivizes underperformance. To neutralize the effects of asking a harder set of questions first (multiplication vs. addition) we controlled the ordering of the task such that half of the workers received addition first, and half multiplication first. Similarly, we also emphasized which arithmetic operator should be performed for half of the workers by making the verb multiply or add (first word of bullet point 1 in Figure 2) bold. To prevent workers from commencing tasks prior to reading the instructions that change only subtly between each half of task they must have acknowledged having read and understood the instructions to reveal the image pairs.

2) AUTOMATED FEEDBACK

The automated feedback system for the second experiment is a variable of interest that represents the quality control scenario. Here, we applied three levels of quality control-based feedback corresponding to three QA treatment groups: 1) feedback disabled (control group) to identify a baseline of quality; 2) automated feedback enabled only in the first part of the task, with it disabled in the second part (initial feedback group); 3) automated feedback enabled only in the second half of the task, with it disabled in the first half (final feedback group).

TABLE 1. Distribution of contributors in the first experiment across 15 conditions.

	None	Fake	Intro	Auto	Wizard
Semantic	17	19	17	18	19
Question	19	17	16	19	18
Translation	16	17	18	19	20

Upon completing a micro-task, workers received standardized feedback responses: "Response recorded" when feedback was disabled. When feedback was enabled a correct solution would reveal "Your answer is fine", and an incorrect solution "Other workers have disagreed with your response". Where the latter response aims to indicate that the answer was not known a priori. To further increase the potential effects of quality control-based feedback, workers were not able to edit their answer once it was committed to the system, thus encouraging later solutions to be cognizant of any feedback received.

We classified responses as either correct or incorrect, resulting in dichotomous quality representation. We refrained from notions of partial correctness in this experiment, as firstly, this is captured in the first experiment, and secondly it is difficult to define a meaningful representation of partial quality without additional contextual information, such as whether the worker misread the image, performed the wrong arithmetic operation, inadvertently struck the wrong key or pressed enter too early etc. vs. having insufficient interest in providing a valid answer. Yet, two aspects are consistent among these examples: (un)intentional human error, and due care and attention to detail, which the provision of feedback will highlight to the worker. Many of these scenarios can also be accommodated in the analysis of the experimental data.

3) MEASUREMENTS

We considered three independent variables: the feedback scenario (control, initial feedback, and final feedback), whether the instruction verb (add/multiply) is bold or not, and whether the worker started with addition or multiplication as well as one dependent variable: mean response quality.

IV. EXPERIMENTAL PROCEDURES

In all conditions for the first experiment, contributors were shown three examples of correctly solved tasks and a description of the task, in the second, only a task description was shown. Table 1 shows the distribution of our contributors by level of quality control method and task complexity for the first experiment. Table 2 shows the distribution of our contributors by treatment group, whether the key instruction verb was bold and whether the task started with addition or multiplication

A. EXPERIMENT 1 WORD-BASED SEMANTIC SIMILARITY

Humans are better than algorithms at rating semantic similarity between two words [6]. Semantic similarity plays

TABLE 2. **Distribution of contributors in the second experiment over all 12 conditions (in parentheses the number of image pairs evaluated).**

	Start with Addition		Start with Multiplication	
	Not Bold	Bold	Not Bold	Bold
Control	42 (1,680)	40 (1,600)	41 (1,640)	41 (1,640)
Initial Feedback	41 (1,640)	41 (1,640)	41 (1,640)	40 (1,600)
Final Feedback	41 (1,640)	42 (1,680)	42 (1,680)	40 (1,600)

an important role for many natural language processing tasks, especially word sense disambiguation and information retrieval [69], [70]. Different algorithmic approaches do exist [71], [72], [73] but are not yet able to reproduce human level results [74]. Involving paid online contributors can reduce costs for such tasks, but consistent response quality is harder to assure than for algorithmic solutions because constantly under-performing contributors remain an issue for such tasks [11].

The task issued in the semantic similarity treatment is itself not very complex, only requiring a good command of English. To ensure this, we restricted contributor's origin to be in the US, UK, or Canada. We further restricted the task using a standard dataset [75] consisting of 353 word pairs. In experiment 1, we recruited 90 contributors and collected ∼9,500 responses on the 353 word pairs.

1) QUESTION ANSWERING

Understanding natural language is still a challenging field for artificial systems [76]. Answering questions given in natural language or finding relevant search results to these questions are, despite the recent success of systems such as IBM Watson [77], unsolved challenges [78], [79]. As standard-use datasets for question answering are corruptible for human annotators with access to the internet, we designed a set of 50 questions so that using the question as a search string will not reveal the correct answer right away.

We randomly selected 10 questions to be test questions for conditions with an introductory test (Intro, Auto, Wizard). We designed sets of possible answers to these 10 test questions by hand. Each answer set had ∼10 answers from at least three different people. Answers were collected offline from students and members of our research group. The response quality of a contributor was estimated by the semantic similarity between the contributor's response and our exemplary answers. We took the highest similarity value as an estimate of quality. The method is calibrated by testing each of the handmade answers against the remaining answers in each set. The average similarity of answers on a scale from 0.0 (no similarity) to 1.0 (perfect similarity) was 0.65 (SD: 0.25). Responses within a margin of one standard deviation were considered acceptable.

Each contributor could answer up to 80 questions. We collected 5,089 responses (57 on average) from 89 contributors on Crowd Flower. We collected 1,017 responses on average for each control level.

2) TEXT TRANSLATION

Text translation is a demanding task even for humans as in-depth knowledge of two different domains, the target and the source language, is required. Various approaches exist; applying crowdsourcing to translation targeted paraphrasing [80] and iterative collaboration between monolingual users [81] are two examples. Other common approaches utilize mono- or bilingual speakers to proofread and correct machine translation results [82]. For our experiment, we use respectively a popular Wikipedia article in German and Vietnamese. Native speakers of German and Vietnamese prepared a set of sentences from this article. For the set, we took the first 150 sentences from the respective article. Headlines, incomplete sentences, and sentences that contained words in a strong dialect were removed. We requested translations for the remaining sentences from contributors via Crowd Flower. As the target language was English, we used the same quality prediction method for conditions that included a pre-test as for the question answering task. Each contributor could translate up to 100 sentences. We collected 2,119 translations for the Vietnamese set and 2,002 translations for the German set (total 4,121) from 90 contributors (46 on average). We collected 825 sentences on average in each control condition.

B. EXPERIMENT 2 IMAGE PROCESSING

As high-capacity supervised machine learning methods have emerged (most notably the advent of deep neural networks) the ability for researchers to handle complex (unstructured) image and video data has significantly accelerated the state-of-the-art in image recognition [83]. Yet, modern models require extremely large (often human labelled) datasets for training [84]. Even with the development of transfer learning [85] here a dataset potentially from a different domain entirely is used to build and preconfigure an initial machine learning model as form of a model bootstrapping, and the existence of many platforms and repositories for labelled (image) data (e.g. [86], openml.org etc.) many researchers still need to resort to some amount of crowd-coding for their domain [87]. Where a prime example is fine-grained recognition (e.g. distinguishing between breeds animals, i.e. categorising dog breeds as opposed to classifying dogs in general [88].

In this experiment, we recruited 492 contributors each recognizing and adding the values of 20 pairs of images and multiplying a different 20 pairs. Thus, each contributor sees 40 distinct image pairs of the 50 we selected for the experiment resulting in 19,680 responses (Table 2 shows the break-down across the 12 conditions).

V. RESULTS

Before we can contextualize results, we must first establish that indeed task complexity influences response quality and that we measure response quality reliably.

TABLE 3. Results of the self-assessment; it is not possible to calculate SD as Crowd Flower only offers aggregated data. From left to right the columns refer to overall satisfaction, instruction clarity, test question fairness, payment, and ease of job.

		Satisfaction	Clarity	fairness	Payment
Similarity	3.8	3.8	3.7	4.5	4.3
Question	3.6	3.4	3.5	4.1	3.7
Translate	3.7	3.9	3.3	4.4	3.1

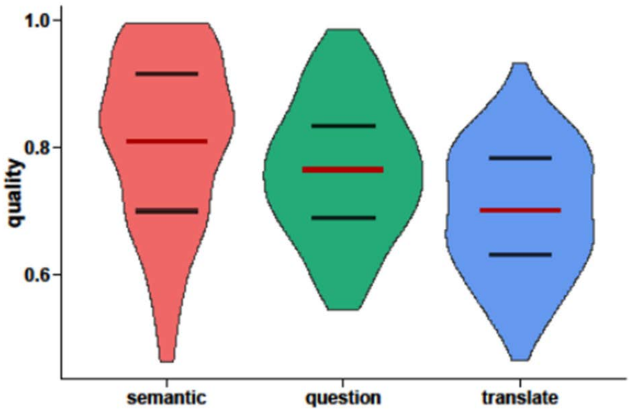


FIGURE 3. Task complexity affects response quality. The most complex task text translation (right) has a significantly lower average response quality than the more simplistic semantic similarity task (left) and the question answering task (middle). The figure shows a violin plot combining a boxplot and a kernel density plot. Thick dark lines indicate 1st and 3rd quartiles the red lines population means.

A. TASK COMPLEXITY AFFECTS RESPONSE QUALITY

We analyse effects for each level of the task complexity factor, assuming that the average response quality deteriorates with higher complexity tasks. As seen in Table 3 and Figure 3 this assumption holds. Although this may seem obvious, it substantiates the initial assumption on task complexity. The Pearson m Product-Moment correlation is 1.0 with an associated $p < 0.001$. This is in line with the self-assessment of contributors' satisfaction survey. We found that Ease of Job negatively correlates with our presumed complexity ranking. Table 3 shows the results of the satisfaction survey: from left to right the columns refer to overall satisfaction, instruction clearness, test question fairness, payment, and ease of job. It is not possible to calculate a SD as Crowd Flower only offers aggregated data.

B. JUDGES AGREE ON QUALITY

Then we ensure that our metric is reasonable. Perceived quality is used as this measure allows investigating quality over different tasks. Table 4 shows that our judges have a substantial agreement on quality throughout all tasks.

Before testing our results for significance, we ensured that our data is suitable for parametric tests. We used the Shapiro-Wilk test for normality [91] for each condition and did not find significant differences from a normal distribution.

TABLE 4. Inter-rater agreement on perceived response quality. The results are homogenous for all three tasks and indicate a substantial agreement between our judges.

	Participants	Judges	Krippendorff's α
Similarity	90	2	0.808
Question	89	2	0.838
Translate	90	2	0.815

TABLE 5. Anova results of main and interaction effects. The first row shows the effect of the quality control method. The second effect of the task. The third their interaction effect.

	df	SS	MS	F	p	sig.
(C)ontrol	4	1.036	0.259	28.988	0.001	***
(T)ask	2	0.557	0.279	31.165	0.001	***
CxT	8	0.220	0.028	3.082	0.002	**
Residuals	254	2.270	0.009			

TABLE 6. Welch two sample t-tests with Holm correction comparing all levels of the quality control factor.

Comp.	M1	SD1	M2	SD2	T	df	p	Sig.
none fake	0.63	0.09	0.80	0.11	-8.21	100	0.00	***
none intro	...	...	0.79	0.12	-7.72	97	0.00	***
none auto	...	...	0.78	0.13	-7.67	105	0.00	***
none wiz.	...	...	0.79	0.13	-8.17	106	0.00	***
fake intro	0.80	0.11	0.79	0.12	0.44	102	0.66	
fake auto	...	...	0.78	0.13	0.74	106	0.46	
fake wiz.	...	...	0.79	0.13	0.25	107	0.80	
intro auto	0.79	0.11	0.78	0.11	0.29	104	0.77	
intro wiz.	...	...	0.79	0.13	-0.20	105	0.85	
auto wiz.	0.78	0.11	...	...	-0.50	111	0.62	

C. QUALITY CONTROL AND TASK COMPLEXITY INTERACT

As we have different numbers of contributors in our conditions, we also verified that our conditions have equal variance for the dependent variable prior to executing an analysis of variance (ANOVA). As the distributions do not differ significantly from normal distributions we use Bartlett's test for homoscedasticity (equal variance) [89]. We found that the variance does not differ significantly between our conditions $t(4) = 2.764$, $p = 0.598$. As our data does not hold evidence that it violates the assumptions of the ANOVA, we analyse main and interaction effects with a two-way ANOVA to compare the effect of quality control and task complexity on the independent variable perceived response quality. Table 5 shows these results.

From the ANOVA results, we conclude that task complexity as well as the used quality control method have a significant influence on the perceived response quality.

TABLE 7. Results of Welch two sample t-tests with Holm correction. Line 1 compares level semantic to level question of the task complexity factor. Line 2 compares level semantic translation and the line three question to translation.

Comp.	M1	SD1	M2	SD2	T	*df*	*p*	Sig.
Sem. Quest.	0.81	0.13	0.77	0.11	2.45	169	0.02	*
Sem. Trans.	...	...	0.70	0.10	6.07	167	0.00	***
Quest Trans.	0.77	0.11	0.70	0.10	4.10	177	0.00	***

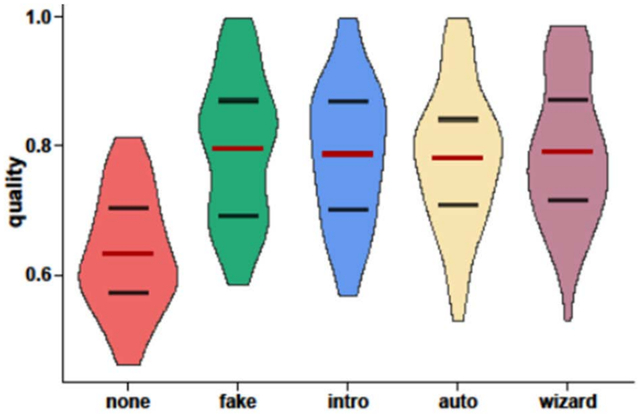


FIGURE 4. Quality control affects response quality only if there is no quality control at all. The differences in means between quality control methods are not significant.

TABLE 8. Means and standard deviations for perceived quality. Rows contain the five different quality control methods and columns the different tasks.

	Semantic		Question		Translation	
	M	SD	M	SD	M	SD
none	0.62	0.09	0.68	0.09	0.60	0.08
intro	0.84	0.11	0.78	0.09	0.74	0.11
fake	0.85	0.10	0.81	0.11	0.72	0.09
auto	0.89	0.07	0.76	0.06	0.70	0.10
wizard	0.83	0.11	0.81	0.13	0.73	0.07

Furthermore, we found a significant interaction between both factors. We use Welch Two Sample t-test with Holm-Bonferroni correction as our post hoc comparison method. Table 6 presents differences in levels of the control factor.

D. ONLY COMPLETE ABSENCE OF QUALITY CONTROL AFFECTS RESPONSE QUALITY

The results indicate that there is a significant difference between the levels ‘‘none of control and the other four levels. The resulting p-values are below the 0.001 alpha-level as seen in Table 7. Other levels do not differ significantly. Table 8 shows means and standard deviations between all levels of our two factors. Figure 4 further illustrates that the finding is constant for all tested tasks.

We also investigated the proportion of constantly underperforming contributors (a contributor below a quality level of 0.6). We found that in all no-quality control conditions we had a substantial number of contributors ($N = 22$) with an average response quality below 0.6. In all other conditions combined, we found 11 contributors under this threshold. The proportion of underperforming contributors in the none conditions is 0.42. Compared to the other conditions with a proportion of only 0.05 this is value is extremely high [68].

In the auto level of the quality control factor, an ML-System predicted the response quality of contributors based on two features (number of characters typed and time needed to complete a request). To estimate the quality of this prediction we calculated the correlation between our ML-systems prediction and the average perceived quality. The ML-system rated responses on a scale with three ordered values (unacceptable (1); acceptable (2); good (3)). As this scale is ordinal and violates the assumptions of Pearson’s Product-Moment correlation we analysed the correlation using Spearman’s ρ. We found a substantial correlation between the predictions and the average perceived quality of our human judges ρ (937020) $= 0.71$, $p < 0.001$. The correlation between the two human judges in comparison is ρ (463061) $= 0.85$, $p<0.001$. In contrast, the human raters who replaced the ML-system in our wizard condition achieved a correlation of ρ (705574) $= 0.78$, $p<0.001$.

E. QUALITY IS ALMOST INDEPENDENT OF THE TYPE OF QUALITY CONTROL USED

In our second experiment we investigated three main effects 1) quality control through the treatment variable (QA Treatment), the task itself either addition or multiplication (add/multiply) and increase in attention through bolding action words in the task description (Bold). We also investigated possible interaction effects between the significant effects. The QA Treatment variable does not show a significant overall impact on the data set (see Table 9). The Task variable encoded which task the user executed either addition or multiplication had a significant effect as well as the bolding of the verbs (add/multiply) in the task description. Finally, showing the addition task before the multiplication task (Addition first) also influenced the quality outcome. Table 9 shows an analysis of variance to test for potentially interesting effects and interactions.

As in the first experiment, the second experiment (image recognition) again shows only minimal non-significant quality differences between the three different quality control conditions (QA Treatment). Table 10 shows the results of our linear model for the three conditions.

A strong contributor to response quality was the task itself. Contributor performance was significantly lower when completing the addition task compared to the multiplication. The reasons for this effect will be discussed in the conclusion section in detail, but the primary reason was the (provoked) misunderstanding of the task description [37], [38]. The addition task has a ~7% higher error rate than the multiplication task (see Figure 5 and Table 11, which illustrated

TABLE 9. The QA Treatment variable does not show a significant overall impact on the data set. The Task variable encoded which task the user executed either addition or multiplication had a significant effect as well as the bolding of the verbs (add/multiply) in the task description. Finally, showing the addition task before the multiplication task (Addition first) also influenced the quality outcome.

	Df	Sum Sq	Mean Sq	F Value	*Pr (>F)*	Sig
QA Treatment	2	0.08	0.042	0.485	0.615	
Task (+ or *)	1	1.70	1.701	19.624	<0.001	***
Bold	1	0.86	0.862	9.953	0.002	**
Addition first	1	0.53	0.526	6.071	0.013	*
Task: Bold	1	0.00	0.004	0.057	0.811	
Bold: Addition first	1	0.01	0.009	0.104	0.747	

TABLE 10. Feedback disabled (control group/intercept), QA Treatment 1) automated feedback enabled only in the first half of the task QA Treatment 2) automated feedback enabled only in the second half of the task.

	Estimate	Error	t-value	F-Value	Sig
Intercept (no in task feedback)	0.853	0.015	55.858	<0.001	***
QA Treatment 1 feedback first	0.015	0.021	0.703	0.482	
QA Treatment 2 feedback second	0.020	0.021	0.933	0.351	

TABLE 11. The addition task shows significantly lower response quality. The reason is a misinterpretation of the term "add" in the task description. Contributors were putting both numbers in sequence instead of adding the number. A 2 and 0 would be interpreted as 20 rather than 2.

	Estimate	Error	t-value	F-Value	Sig
Intercept	0.903	0.012	73.057	<0.001	***
Addition	-0.077	0.017	-4.404	<0.001	***

that addition tasks are incorrect more often, but multiplication tasks have a higher degree of error). It's interesting to note that when comparing the individual task success rates, i.e., proportion of contributors correctly solving the addition task versus the multiplication task on the same image pair the success rates are not correlated (control: $\rho = -0.019$, p = 0.897; initial feedback: $\rho = -0.174$ p = 0.008; final feedback: $\rho = 0.141$ p = 0.326; all treatment groups: $\rho = 0.072$, p = 0.380). This further suggests a misunderstanding of the task, and not an issue of contributors recognizing the image pair. However, the mean degree with which the contributor is incorrect is slightly positively correlated when feedback is provided, and in general across all treatment groups, but not in the control group (control: $\rho = 0.114$, p = 0.428; initial feed-back: $\rho = 0.273$ p = 0.056; final feedback: $\rho = 0.241$ p = 0.092; all treatment groups: $\rho = 0.215$, p = 0.008).

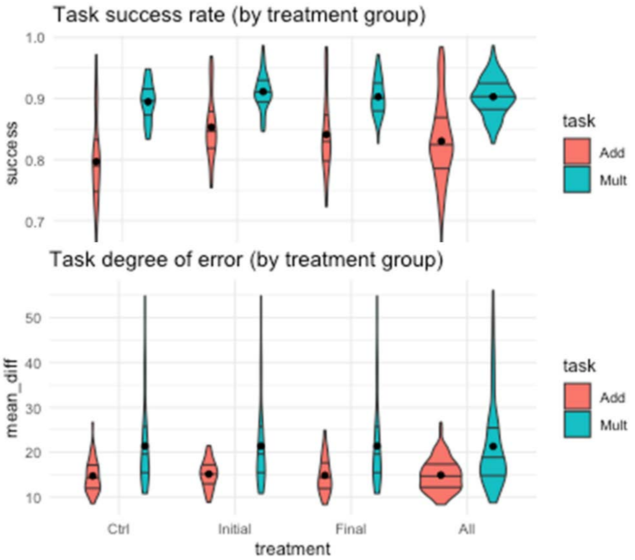


FIGURE 5. Violin plots illustrating Task Success Rate (top) and Degree of Error (bottom) by treatment group. In each violin, the black circle indicates the mean.

TABLE 12. Whether the contributor was asked to complete the addition, or the multiplication task first had a significant impact on the overall response quality of a contributor. The misconception from the addition task seems to carry over to the multiplication task in some cases.

	Estimate	Error	t-value	F-Value	Sig
Intercept	0.886	0.012	71.176	<0.001	***
Addition first	-0.042	0.017	-2.408	0.016	*

TABLE 13. The bolding of the verbs (add/multiply) in the task description had a significantly positive impact on worker performance.

	Estimate	Error	t-value	F-Value	Sig
Intercept	0.838	0.012	71.176	<0.001	***
Bold	0.054	0.017	-2.408	0.016	*

This observation is related to another significant effect in the data. If the addition task is shown as the first task group, the negative effect from the wording is carried through to the multiplication task. The overall quality is reduced by ∼4% when the addition task is shown first. This can also be seen in Figure 6 (top), where when addition is shown first the success rate of addition tasks is lower. Conversely, this is not present in multiplication tasks. Table 12 displays the results of this analysis.

F. HIGHLIGHTING ACTION WORDS SIGNIFICANTLY INCREASES RESPONSE QUALITY

Using typeset Bold on the verbs (add/multiply) in the task description did increase response quality for both task types (see Figure 6 (bottom)). It also increased the response quality equally for the order of tasks. The carried negative effect of the addition was mitigated by the bolding of verbs. The bolding increases the average performance by >6%. The increase in quality is consistent across all other variables and can be observed with almost the same effect size in all QA

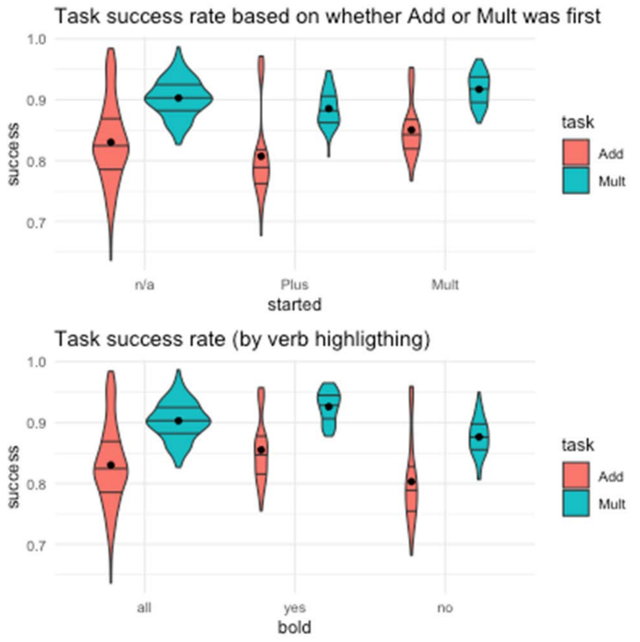


FIGURE 6. **Violin plots illustrating Task success rates according to whether addition or multiplication tasks were first (top) and whether keywords "Add" and "Multiple" were bold in the instructions or not (bottom). In both plots, the left two violins illustrate the general distribution for Add and Mult, and the black circle, the mean.**

Treatment conditions and across all other factors. Table 13 illustrates these results.

VI. DISCUSSION AND CONCLUSION

In this article, we have explored two experimental scenarios to shed light on the relationships between quality control, elements of task design, and response quality structured as two research questions: *RQ1: what characterisations of response quality can be linked to different quality control mechanisms?* and *RQ2: what impacts of task design have larger effects on response quality then different mechanisms of quality control?*

The first experiment sought to highlight the impact of varying degrees of sophistication in the design of the quality control mechanism vs. the complexity of the task to be performed. We saw that more complicated tasks (text translation) were not in need of more complex (e.g., human-based or machine learning-based) quality control mechanisms. In fact, we observed no statistically significant improvement in response quality across the quality control mechanisms applied. We did, however, observe a structural difference in response quality between no quality control and *some kind* of quality control. Even in the presence of "faking" the quality control, i.e., announcing quality control mechanisms, but in fact doing nothing the response quality and number of underperforming contributors improved.

Our second experiment sought to build on and refine these observations. We contrasted the effects of quality control methods with the effects of subtle changes in the task description and ordering of tasks. We again observed that the presence of a feedback-based quality control mechanism increases output quality in the image recognition and reasoning tasks. We also observed the effects of very subtle changes in the task description (boldening parts of the description (add/multiply) or the ordering of the two different tasks (addition first vs. multiplication first). We found that these subtle changes in the task description and presentation have more impact on response quality (7% and 4%) than the absence of a control-based QA-Treatment (2% and 1.5%). This illustrates that it is more effective to interact with constantly underperforming contributors to understand the reason for their actions rather than treating them as mere computational elements; otherwise said, to support their intrinsic motivation rather than enforce extrinsic motivation. The goal of quality assurance measures should foremost be to understand possible misconceptions rather than control of contributors. The effects introduced by poor task design and task descriptions do outweigh the impact of so-called "cheaters". In contrast, interacting with these underperforming contributors can enhance quality and satisfaction on both sides.

Returning to RQ1, we observed that only minimal non-significant differences between different quality control mechanisms on response quality. We observed this in both experiments, which capture a wide range of crowd tasks (in NLP, and image recognition / processing). Similarly, we observed that the number of underperforming workers increases in the absence of any quality control announcement. This is not surprising, however, in the case where quality control is announced, but not performed, there was also minimal non-significant differences to technically advanced mechanisms of quality control.

For RQ2, we can (perhaps not surprisingly) note that task complexity has an impact on response quality. Yet, it is surprising that increasing the level of sophistication in the quality control mechanism for more complex tasks is less impactful than the increase in task complexity itself. We also observed that contributor performance was tightly linked to simple design aspects of the task: making key words bold, task ordering, and a small (yet still significant) impact based upon when in the task quality control feedback occurs; we observed a slightly reduced rate of error in tasks when feedback was provided earlier in the task design.

From these findings, we propose the following suggestions on how practitioners and the research domain can apply quality control to reduce underperforming contributors with a goal towards increasing response quality:

1) **Mention quality control:** The mention of a required (qualification) test or similarly appropriate mechanism (i.e., the fake level in experiment 1) is sufficient to deter "poor" contributors. Using this alone, we observed an increase of more than 25% in response quality.
2) **Keep quality control simple and practical:** Implementing basic quality control and feedback is sufficient to foster diligent work. We would not advocate only faking the quality control approach: contributors would

realise, as they often share task information amongst themselves [41], [46]. As a pragmatic solution, we have shown that machine learning methods can provide a suitably automated method for quality control (in line with [79]). Or in other words, ML can be leveraged to predict response quality on the fly.

3) **Focus on a clear task description:** Significant care is needed to make sure that key elements of the task are very clear [37], [38]. We observed how just bolding specific parts of the task description can have a more significant effect on response quality than implementing quality control mechanisms. This could be as simple as directing focus of the contributor to key parts of task description [90].
4) **Consider contributor training:** whilst not explored in this article, it would be worthwhile to investigate tangible incentives for contributors to upskill towards improving their ability to reliably contribute to additional or high(er) complexity tasks (see e.g., [91], [92], [93], [94], [95]).

In terms of training, we argue that the most basic way to promote this in task learning is the interaction between contributors and requesters. Rather than seeing underperforming contributors as a nuisance, they might very well be a valuable contributor who is acting diligently yet regardless underperforming. As we have demonstrated, even small and subtle changes in the task description can have a more dominant impact on response quality than even sophisticated control-based QA methods. Yet, even so we also know that it is harder to achieve high response quality in high complexity tasks. Thus, our suggestion is that instead of investing in complex, resource demand mechanisms for quality control (this is not a dismissal of research into mechanisms for quality control), we should rather seek to develop approaches to improve contributor training and skill development to globally improve quality [1], [8], [17], [18], [90], [96]. One possibility here is to establish a means to certify training and development for crowd contributors. Not only would this foster better contributors over time, but also help securing contributors' rights in crowd labour markets [1], [8], [18]. Thus, key future work building on our results should focus on how to operationalise viable training regimes and appropriately incentivise contributors to engage with such programmes.

REFERENCES

[1] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 76–81, Mar./Apr. 2013, doi: 10.1109/MIC.2013.20.

[2] M. Krause, F. M. Afzali, S. Caton, and M. Hall, "Is quality control pointless?" in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 5279–5288, doi: 10.24251/hicss.2019.636.

[3] D. Spurk and C. Straub, "Flexible employment relationships and careers in times of the COVID-19 pandemic," *J. Vocational Behav.*, vol. 119, Jun. 2020, Art. no. 103435, doi: 10.1016/j.jvb.2020.103435.

[4] A. Desai, J. Warner, N. Kuderer, M. Thompson, C. Painter, G. Lyman, and G. Lopes, "Crowdsourcing a crisis response for COVID-19 in oncology," *Nature Cancer*, vol. 1, no. 5, pp. 473–476, May 2020, doi: 10.1038/s43018-020-0065-z.

[5] X. Deng and K. D. Joshi, "Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions," *J. Assoc. Inf. Syst.*, vol. 17, no. 10, pp. 648–673, Nov. 2016, doi: 10.17705/1jais.00441.

[6] N. Batram, M. Krause, and P. O. Dehaye, "Comparing human and algorithm performance on estimating word-based semantic similarity," in *Proc. Int. Conf. Social Inform.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 8852, 2014, pp. 452–460, doi: 10.1007/978-3-319-15168-7_55.

[7] C. Sarasua and M. Thimm, "Microtask available, send us your CV," in *Proc. Int. Conf. Cloud Green Comput.*, Sep. 2013, pp. 521–524, doi: 10.1109/CGC.2013.87.

[8] C. Dukat and S. Caton, "Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities," in *Proc. Int. Conf. Cloud Green Comput.*, Sep. 2013, pp. 536–540, doi: 10.1109/CGC.2013.90.

[9] J. Wang, P. Ipeirotis, and F. Provost, "Quality-based pricing for crowdsourced workers," NYU Stern Res., Work. Paper CBA-13-06, 2013, pp. 1–46. [Online]. Available: https://ipeirotis.org/wpcontent/uploads/2013/06/CBA-13-06.pdf

[10] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon mechanical Turk," in *Proc. ACM SIGKDD Workshop Hum. Comput. (HCOMP)*, 2010, pp. 1–3.

[11] M. Krause and R. Porzel, "It is about time: Time aware quality management for interactive systems with humans in the loop," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, 2013, p. 163, doi: 10.1145/2468356.2468386.

[12] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing," in *Proc. AAAI Workshop Hum. Comput.*, 2011, pp. 43–48.

[13] M. Lease, "On quality control and machine learning in crowdsourcing," in *Proc. AAAI Workshop*, vol. 11, 2011, pp. 97–102. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-80055052639&partnerID=40&md5=86ee90fbcf4ed7ba86f4c810775ad194

[14] Q. Chen, J. Bragg, L. B. Chilton, and D. S. Weld, "Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, p. 14, doi: 10.1145/3290605.3300761.

[15] K. Hata, R. Krishna, L. Fei-Fei, and M. S. Bernstein, "A glimpse far into the future," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2017, pp. 889–901, doi: 10.1145/2998181.2998248.

[16] X. Deng, K. D. D. Joshi, and R. D. Galliers, "The duality of empowerment and marginalization in microtask crowdsourcing," *MIS Quart.*, vol. 40, pp. 1–24, Jun. 2016.

[17] K. Zyskowski, M. R. Morris, J. P. Bigham, M. L. Gray, and S. K. Kane, "Accessible crowdwork: Understanding the value in and challenge of microtask employment for people with disabilities," in *Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2015, pp. 1682–1693. [Online]. Available: http://dl.acm.org/citation.cfm?id=2675133.2675158&coll=DL&dl=ACM&CFID=674687730&CFTOKEN=91916307

[18] S. C. Kingsley, M. L. Gray, and S. Suri, "Accounting for market frictions and power asymmetries in online labor markets," *Policy Internet*, vol. 7, no. 4, pp. 383–400, Dec. 2015, doi: 10.1002/poi3.111.

[19] U. Gadiraju and N. Gupta, "Dealing with sub-optimal crowd work: Implications of current quality control practices," in *Proc. CHI*, 2016, pp. 15–20.

[20] R. Kazman and H.-M. Chen, "The metropolis model a new logic for development of crowdsourced systems," *Commun. ACM*, vol. 52, no. 7, pp. 76–84, Jul. 2009, doi: 10.1145/1538788.1538808.

[21] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: Using implicit behavioral measures to predict task performance," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2011, pp. 13–22, doi: 10.1145/2047196.2047199.

[22] A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino, "Keep it simple: Reward and task design in crowdsourcing," in *Proc. Biannual Conf. Italian Chapter SIGCHI*, 2013, p. 14.

[23] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms," in *Proc. CEUR Workshop*, vol. 842, 2012, pp. 20–25.

[24] E. Newell and D. Ruths, "How one microtask affects another," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 3155–3166, doi: 10.1145/2858036.2858490.

[25] C. J. Cai, S. T. Iqbal, and J. Teevan, ''Chain reactions: The impact of order on microtask chains,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 3143–3154, doi: 10.1145/2858036.2858237.

[26] D. Chandler and A. Kapelner, ''Breaking monotony with meaning: Motivation in crowdsourcing markets,'' *J. Econ. Behav. Org.*, vol. 90, pp. 123–133, Jun. 2013, doi: 10.1016/j.jebo.2013.03.003.

[27] J. Prpic and P. Shukla, ''Crowd science: Measurements, models, and methods,'' in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 4365–4374, doi: 10.1109/HICSS.2016.542.

[28] C. Gerber and M. Krzywdzinski, ''Brave new digital work? New forms of performance control in crowdwork,'' in *Research in the Sociology of Work*, vol. 33. Bingley, U.K.: Emerald Group Publishing Ltd., 2019, pp. 121–143, doi: 10.1108/S0277-283320190000033008.

[29] J. Oppenlaender, K. Milland, A. Visuri, P. Ipeirotis, and S. Hosio, ''Creativity on paid crowdsourcing platforms,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–14, doi: 10.1145/3313831.3376677.

[30] G. Smith, R. C. Richards, and J. Gastil, ''The potential of participedia as a crowdsourcing tool for comparative analysis of democratic innovations,'' *Policy Internet*, vol. 7, no. 2, pp. 243–262, Jun. 2015.

[31] J. Prpić, A. Taeihagh, and J. Melton, ''The fundamentals of policy crowdsourcing,'' *Policy Internet*, vol. 7, no. 3, pp. 340–361, Sep. 2015, doi: 10.1002/poi3.102.

[32] C. Niemeyer, T. Teubner, M. Hall, and C. Weinhardt, ''The impact of dynamic feedback and personal budgets on arousal and funding behaviour in participatory budgeting,'' *Group Decis. Negotiation*, vol. 27, no. 4, pp. 611–636, Aug. 2018, doi: 10.1007/s10726-018-9578-6.

[33] D. Friess and C. Eilders, ''A systematic review of online deliberation research,'' *Policy Internet*, vol. 7, no. 3, pp. 319–339, Sep. 2015.

[34] E. Law, K. Z. Gajos, A. Wiggins, M. L. Gray, and A. Williams, ''Crowdsourcing as a tool for research: Implications of uncertainty,'' in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2017, pp. 1544–1561, doi: 10.1145/2998181.2998197.

[35] G. Kazai, J. Kamps, and N. Milic-Frayling, ''The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy,'' in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 3–6.

[36] G. Kazai, J. Kamps, and N. Milic-Frayling, ''Worker types and personality traits in crowdsourcing relevance labels,'' in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, p. 1941, doi: 10.1145/2063576.2063860.

[37] U. Gadiraju, J. Yang, and A. Bozzon, ''Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing,'' in *Proc. 28th ACM Conf. Hypertext Social Media*, Jul. 2017, pp. 5–14, doi: 10.1145/3078714.3078715.

[38] J. Yang, J. Redi, G. Demartini, and A. Bozzon, ''Modeling task complexity in crowdsourcing,'' in *Proc. 4th AAAI Conf. Hum. Comput. Crowdsourcing*, Oct. 2016, pp. 249–258.

[39] H. Corrigan-Gibbs, N. Gupta, C. Northcutt, E. Cutrell, and W. Thies, ''Deterring cheating in online environments,'' *ACM Trans. Comput.-Hum. Interact.*, vol. 22, no. 6, pp. 1–23, Dec. 2015, doi: 10.1145/2810239.

[40] A. Kittur, E. H. Chi, and B. Suh, ''Crowdsourcing user studies with mechanical Turk,'' in *Proc. 26th Annu. CHI Conf. Hum. Factors Comput. Syst.*, 2008, pp. 453–456, doi: 10.1145/1357054.1357127.

[41] M. L. Gray, S. Suri, S. S. Ali, and D. Kulkarni, ''The crowd is a collaborative network,'' in *Proc. 19th ACM Conf. Comput.-Supported Cooperat. Work Social Comput.*, Feb. 2016, pp. 134–147, doi: 10.1145/2818048.2819942.

[42] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, ''The future of crowd work,'' in *Proc. Conf. Comput. Supported Cooperat. Work*, 2013, p. 1301, doi: 10.1145/2441776.2441923.

[43] D. Geiger, S. Seedorf, R. Nickerson, and M. Schader, ''Managing the crowd: Towards a taxonomy of crowdsourcing processes,'' in *Proc. 17th Americas Conf. Inf. Syst.*, 2011, pp. 1–11, doi: 10.1113/jphysiol.2003.045575.

[44] U. Gadiraju, G. Demartini, R. Kawase, and S. Dietze, ''Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection,'' *Comput. Supported Cooperat. Work*, vol. 28, no. 5, pp. 815–841, Sep. 2019, doi: 10.1007/s10606-018-9336-y.

[45] U. Gadiraju, B. Fetahu, R. Kawase, P. Siehndel, and S. Dietze, ''Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks,'' *ACM Trans. Comput.-Hum. Interact.*, vol. 24, no. 4, pp. 1–26, Sep. 2017, doi: 10.1145/3119930.

[46] M. Yin, M. L. Gray, S. Suri, and J. W. Vaughan, ''The communication network within the crowd,'' in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 1293–1303, doi: 10.1145/2872427.2883036.

[47] W.-C. Chen, S. Suri, and M. L. Gray, ''More than money: Correlation among worker demographics, motivations, and participation in online labor market,'' in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, Jul. 2019, pp. 134–145, Accessed: Oct. 26, 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3216

[48] J. J. Chen, N. J. Menezes, A. D. Bradley, and T. A. North, ''Opportunities for crowdsourcing research on Amazon mechanical Turk,'' *Interface*, vol. 5, no. 3, pp. 1–4, 2011, doi: 10.1145/1357054.1357127.

[49] A. J. Berinsky, G. A. Huber, and G. S. Lenz, ''Evaluating online labor markets for experimental research: Amazon.com's mechanical Turk,'' *Political Anal.*, vol. 20, no. 3, pp. 351–368, 2012, doi: 10.1093/pan/mpr057.

[50] V. S. Sheng, F. Provost, and P. G. Ipeirotis, ''Get another label? Improving data quality and data mining using multiple, noisy labelers,'' in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, p. 614, doi: 10.1145/1401890.1401965.

[51] R. Kern, H. Thies, and G. Satzger, ''Statistical quality control for human-based electronic services,'' in *Proc. Int. Conf. Service-Oriented Comput.*, 2010, pp. 1–17.

[52] B. D. Rosso, K. H. Dekas, and A. Wrzesniewski, ''On the meaning of work: A theoretical integration and review,'' *Res. Organizational Behav.*, vol. 30, pp. 91–127, Jan. 2010, doi: 10.1016/j.riob.2010.09.001.

[53] D. Ariely, E. Kamenica, and D. Prelec, ''Man's search for meaning: The case of Legos,'' *J. Econ. Behav. Org.*, vol. 67, nos. 3–4, pp. 671–677, Sep. 2008, doi: 10.1016/j.jebo.2008.01.004.

[54] A. J. Quinn and B. B. Bederson, ''Human computation: A survey and taxonomy of a growing field,'' in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2011, pp. 1403–1412, doi: 10.1145/1978942.1979148.

[55] M. Bernstein, E. H. Chi, L. Chilton, B. Hartmann, A. Kittur, and R. C. Miller, ''Crowdsourcing and human computation: Systems, studies and platforms,'' in *Proc. Annu. Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2011, pp. 3012–3032, doi: 10.1145/1979742.1979593.

[56] W. Ke, C.-H. Tan, C.-L. Sia, and K.-K. Wei, ''Inducing intrinsic motivation to explore the enterprise system: The supremacy of organizational levers,'' *J. Manag. Inf. Syst.*, vol. 29, no. 3, pp. 257–290, Dec. 2012, doi: 10.2753/mis0742-1222290308.

[57] R. M. Ryan and E. L. Deci, ''Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being,'' *Amer. Psychol.*, vol. 55, no. 1, p. 68, 2000, doi: 10.1002/jsfa.2740050407.

[58] L. Zhao, B. Detlor, and C. E. Connelly, ''Sharing knowledge in social Q&A sites: The unintended consequences of extrinsic motivation,'' *J. Manage. Inf. Syst.*, vol. 33, no. 1, pp. 70–100, Jan. 2016, doi: 10.1080/07421222.2016.1172459.

[59] B. Mcinnis, D. Cosley, C. Nam, and G. Leshed, ''Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon mechanical Turk,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 2271–2282, doi: 10.1145/2858036.2858539.

[60] N. Runge, N. Kilian, J. Smeddinck, and M. Krause, ''Predicting crowd-based translation quality with language-independent feature vectors,'' in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012, pp. 114–115.

[61] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[62] D. Cieslak and N. Chawla, ''Learning decision trees for unbalanced data,'' in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2008, pp. 241–256.

[63] L. Breiman, ''Technical note: Some properties of splitting criteria,'' *Mach. Learn.*, vol. 24, no. 1, pp. 41–47, Jul. 1996, doi: 10.1007/BF00117831.

[64] F. Pedregosa and G. Varoquaux, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12 no. 10, pp. 2825–2830, 2012.

[65] A. P. Dempster, N. M. Laird, and D. B. Rubin, ''Maximum likelihood from incomplete data via the *EM* algorithm,'' *J. Roy. Stat. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–22, Sep. 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[66] K. Krippendorff, ''Agreement and information in the reliability of coding,'' *Commun. Methods Measures*, vol. 5, no. 2, pp. 93–112, Apr. 2011, doi: 10.1080/19312458.2011.568376.

[67] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA, USA: SAGE, 2018.

[68] K. Krippendorff, ''Estimating the reliability, systematic error and random error of interval data,'' *Educ. Psychol. Meas.*, vol. 30, no. 1, pp. 61–70, 1970.

[69] J. Feng, Y. Zhou, and T. Martin, "Sentence similarity based on relevance," in *Proc. 12th Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst.*, vol. 2008, pp. 832–839.

[70] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.*, vol. 41, no. 2, p. 10, 2009, doi: 10.1145/1459352.1459355.

[71] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. AAAI*, 2006, pp. 1419–1424, doi: 10.1.1.231.9545.

[72] D. Yang and D. M. W. Powers, "Measuring semantic similarity in the taxonomy of WordNet," in *Proc. 28th Australas. Comput. Sci. Conf.*, 2005, pp. 315–322.

[73] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, pp. 1–6, doi: 10.1.1.55.5277.

[74] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 337–346, doi: 10.1145/1963405.1963455.

[75] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 406–414.

[76] M. Krause, "Designing systems with homo ludens in the loop," in *Handbook of Human Computation*, P. Michelucci and K. Greene, Eds. New York, NY, USA: Springer, 2014, pp. 393–409, doi: 10.1007/978-1-4614-8806-4.

[77] D. Ferrucci, E. Brown, J. Chu-Carroll, and J. Fan, "Building Watson: An overview of the DeepQA project," *AI Mag.*, vol. 31, no. 3, pp. 59–79, 2010.

[78] H. Aras, M. Krause, A. Haller, and R. Malaka, "Webpardy: Harvesting QA by HC," in *Proc. ACM SIGKDD Workshop Hum. Comput.*, 2010, pp. 49–52, doi: 10.1145/1837885.1837900.

[79] M. Krause, *Homo Ludens in the Loop: Playful Human Computation Systems*. Hamburg, Germany: Tredition GmbH, 2014.

[80] P. Resnik, H. Chang, O. Buzek, and B. B. Bederson, "Using monolingual human computation to improve language translation via targeted paraphrase," in *Proc. ACM SIGKDD Workshop Hum. Comput.*, 2010, pp. 1–4.

[81] C. Hu, B. B. Bederson, P. Resnik, and Y. Kronrod, "MonoTrans2: A new human computation system to support monolingual translation," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2011, pp. 2–5.

[82] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proc. ACL*, 2011, pp. 1220–1229.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[84] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[85] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[86] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 487–495.

[87] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[88] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 580–587, doi: 10.1109/CVPR.2013.81.

[89] M. S. Bartlett, "Properties of sufficiency and statistical tests," *Proc. Roy. Soc. London. Ser. A, Math. Phys. Sci.*, vol. 160, no. 901, pp. 268–282, 1937, doi: 10.1098/rspa.1937.0109.

[90] M. Streuer, M. Krause, M. Hall, and S. Dow, "On-the-job learning for micro-task workers," in *Proc. Hum. Comput.*, 2017.

[91] M. Krause, M. Hall, J. J. Williams, P. Paritosh, J. Prip, and S. Caton, "Connecting online work and online education at scale," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, May 2016, pp. 3536–3541, doi: 10.1145/2851581.2856488.

[92] R. Suzuki, N. Salehi, M. S. Lam, J. C. Marroquin, and M. S. Bernstein, "Atelier: Repurposing expert crowdsourcing tasks as micro-internships," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 1–56, doi: 10.1145/2858036.2858121.

[93] M. Hall, M. Friend, and M. Krause, "Micro-internships on the margins," in *Proc. Int. Conf. Universal Access Hum.-Comput. Interact.*, 2018, pp. 486–495.

[94] M. Hall, A. E. Pavlakis, and M. Friend, "Work-learn: Necessary and sufficient conditions for upskilling homeless adults with entry-level programming and tech sector skills," in *Proc. ICIS TREOs*, vol. 53, 2021.

[95] M. Krause, J. Smeddnick, and D. Schioeberg, "Mooqita: Empowering hidden talents with a novel work-learn model," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–10.

[96] J. Hui, E. M. Gerber, L. Dombrowski, M. L. Gray, A. Marcus, and N. Salehi, "Computer-supported career development in the future of work," in *Proc. Companion ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Oct. 2018, pp. 133–136, doi: 10.1145/3272973.3274545.

MARGERET HALL received the Ph.D. degree in social computing from the School of Economics and Industrial Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2015. She is currently an Assistant Professor in strategic business analytics with the Vienna University of Economics and Business, where she is the Founding Co-Director of the Center for Strategic Business Analytics. Her research interests include social computing, image analysis, sentiment analysis, and data multimodality.

MOHAMMAD FARHAD AFZALI received the B.C.S. degree from Kardan University, and the M.S. degree in management information systems and the Ph.D. degree in information technology from the University of Nebraska Omaha (UNO). Besides serving as an Adjunct Instructor at UNO, he is currently working as a Data Scientist outside academia. In 2007, he worked with the International Rescue Committee, as a Database Officer. In 2008, he join the United Nations Assistance Mission in Afghanistan (UNAMA). He worked for UNAMA for six years, as an Application Developer.

MARKUS KRAUSE is currently the Founder and the Manager of the Joint Intelligence Program at Google. He founded and co-founded a diverse portfolio of companies in the Bay Area and was a Postdoctoral Scholar at the UC Berkeley. In his research, he investigates how human an artificial intelligence can form systems that solve problems neither humans nor machines can solve alone. He also investigates the intersection of working and learning at scale. He is involved in projects exploring self-directed, self-organized, and personalized online education at scale. He worked as a Professional Game Designer and the Art-Director. He led the Research Team of the award winning WuppDi game. The game helps Parkinson's disease patients in their daily fight against their affliction.

SIMON CATON received the B.Sc. degree (Hons.) in computer science, in 2005, and the Ph.D. degree in computer science, in 2010. He worked as a Postdoctoral Researcher at the Karlsruhe Institute of Technology, Germany, from 2010 to 2014, and as a Lecturer of data analytics at the National College of Ireland, between 2014 and 2019. He is currently an Assistant Professor with the School of Computer Science, University College Dublin. His research interests include the applications of machine learning across the domains of social media, quantum computing, and parallel and distributed computing.