9-2016

Is Quality Control Pointless?

Markus Karuse
University of California - Berkeley

Margeret A. Hall
University of Nebraska at Omaha, mahall@unomaha.edu

Simon James Caton
National College of Ireland

Is Quality Control Pointless?

Markus Krause,[a] Margeret Hall,[b] Simon Caton[c]

[a]UC Berkeley, ISCI, 1947 Center St. Ste. 600, Berkeley, CA 94704

[b]University of Nebraska Omaha, School of Interdisciplinary Informatics, 1110 S 67th St. Omaha, NE 68182

[c]National College of Ireland, Mayor Street, Dublin 1

public.markus.krause@gmail.com; mahall@unomaha.edu; simon.caton@ncirl.ie

Abstract

Intrinsic to the transition towards, and necessary for the success of digital platforms as a service (at scale) is the notion of human computation. Going beyond 'the wisdom of the crowd', human computation is the engine that powers platforms and services that are now ubiquitous like Duolingo and Wikipedia. In spite of increasing research and population interest, several issues remain open and in debate on large-scale human computation projects. Quality control is first among these discussions. We conducted an experiment with three different tasks of varying complexity and five different methods to distinguish and protect against constantly underperforming contributors. We illustrate that minimal quality control is enough to repel constantly underperforming contributors and that this effect is constant across tasks of varying complexity.

Introduction

At the center of the debate on large-scale human computation projects is invariably a discussion of quality. This is due in part to the fact that a suitable and scalable mechanism for the *ex-ante* detection of constantly underperforming contributors has yet to be introduced. Commonly used tactics include qualification tests (Batram et al. 2014), pre-set qualifications (Sarasua & Thimm 2013); constructing trust models to determine the probability of diligent work (Wang et al. 2013; Ipeirotis et al. 2010; Krause & Porzel 2013); hidden gold standard questions (Oleson et al. 2011); and the use of metrics such as solution acceptance (Dukat & Caton 2013) (Related Work).

In this work, we present a study that illustrates that constantly underperforming contributors will not take on tasks that feature quality control mechanism. Here we note that we do not use the term spammer, as we cannot predict the intention of our contributors. Our research employs a 3 x 5 factorial experimental design of three task types with varying complexities and five different quality control methods to measure the impact of quality control and task complexity on output quality (Study Design).

Our results indicate that the quality control methods does not have a significant impact on response quality. In the experiment, it was sufficient to simply state that a qualification test is necessary to repel constantly underperforming contributors (Results).

In our experiment, most contributors were diligent. Constantly underperforming contributors by our definition were only present in conditions with no quality control. We argue that expansive quality control support and applications are unwarranted and that more resources should be dedicated to adequate training of contributors in order to raise the overall quality of crowdsourced contributions (Conclusion).

Related Work

The aspect of quality and quality control within Crowdsourcing platforms appears in many ways. In general, Kittur et al. (Kittur et al. 2013) differentiate "up-front task design" and "post-hoc result analysis" as the two main approaches to control work quality in the context of crowdsourcing.

Pre-Selecting Contributors

Crowdsourcing platforms provide the means for employers to pre-select contributors based upon specific task requirements or employer preferences. Geiger et al. (Geiger et al. 2011) define pre-selection as "a means of ensuring a minimum ex-ante quality level of contributions." In other words, an employer will use a pre-selection process or test to mitigate the risk of poor quality solutions. Pre-selection screens potential contributors based upon the completion of some process that demonstrates certain knowledge or skills.

Oleson et al. (Oleson et al. 2011) examine this process, which is typically performed via multiple-choice tests, and highlight as well as subsequently criticise a key assumption in this approach: that 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. Similarly, if the contributor fails the test they may be banned from the task but not necessarily for the right reasons. This method, however, is simple to implement and also typically performs well. Pre-selection via qualification tests is also likely to act as a barrier for "scammer" contributors, but diligent contributors may also not select the task due to an increased effort on their part. Answers to a qualification test may also be shared amongst users, which will reduce its effectiveness. A closely related point is the representativeness of the test to the task.

Qualification Tests

Some platforms use a qualification test, to not only determine the abilities of a contributor, but also access and assess their basic properties, as this information is often not available to

crowd employers. Stolee and Elbaum (Stolee & Elbaum 2010) and Chen et al. (Chen et al. 2011) are examples here. They state that a qualification can also capture demographic (and similar) properties of the contributor, for example geographical location. This does, however, massively distort the concept of a qualification if personal attributes of a contributor are considered.

One issue that seems to be overlooked in the literature is the transferability and transitivity of a crowd sourcing qualification. Qualifications can be transferred between tasks of the same employer, due to the knowledge that an employer may have on previous interactions with a contributor. This is not the case for other employers of "similar" tasks. From this point of view, Dukat and Caton argued in (Dukat & Caton 2013) that a lack of standardized testing or at least an accepted definition of certain qualifications presents itself as an urgently needed facet of crowdsourcing platforms, in order to prevent contributors redundantly performing "similar" qualification tests.

In Task Quality Control

An alternative method to assessing contributor quality proposed by Sheng et al. and Ipeirotis et al. (Sheng et al. 2008; Ipeirotis et al. 2010) is to infer a level of trust in the contributor via 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 it is very difficult to measure quantitatively; key for approximating (automatically) a contributor's propensity for a diligent or reliable work. Thus, Kern et al (Kern et al. 2010) capture the trustworthiness of contributor based on prior experience. They redundantly schedule tasks to multiple contributors to provide a basis to compare and estimate contributor reliability. This method, demonstrated to yield high quality solutions. Yet without careful management the method is expensive in terms of redundantly issuing tasks (direct costs) and the additional (computational) effort needed to assess solution quality. Similarly, managing the crowd with respect to "rejected" answers can have other adverse effects, especially if the contributor has acted diligently. Oleson et al. (Oleson et al. 2011) identify possible effects as: a loss in reputation for the employer, directed employee complaints, and the potential black listing of diligent contributors.

Oleson et al. (Oleson et al. 2011) propose the use of gold standard questions to assess solution quality and contributor ability. In their approach, subtasks with known solutions are injected into the task. The presence of these questions enables the accuracy of a given contributor to be estimated in task, and help improve the quality of their solutions by providing an explanation why the solution is incorrect. Therefore, contributors can receive instant feedback on the accuracy of their performance. The approach, however, is limited to tasks that have a finite set of definite answers, and is inappropriate for tasks that rely on forms of subjectivity. However, such a mechanism provides a basis to also train a contributor, and enable a contributor to self-evaluate their performance through system feedback. Where the latter facilitates an integral element in the definition of competence: the evaluation of self-efficacy.

Competence and Self-Efficacy

Bringing in the notion of self-efficacy, and by extension competence, leads to the discussion on what it means to be a competent crowd contributor. Dukat and Caton (Dukat & Caton 2013) as well as Dow et al. (Dow et al. 2012) discuss this at a high level. They identify that a competent crowd contributor is not only able to complete the task, but also willing to undertake the task at hand diligently, and finally cognoscente of their own limitations. We can observe that today crowdsourcing platforms use several mechanisms to assess contributor reliability and capabilities before, during or after task completion. The focus of these measures is arguably to minimize the risk of lower quality solutions, but ultimately do not actually provide concrete quality assurance at the individual contributor level. At best these measure act as an artificial proxy for contributor "competence" and mainly allow employers to pre-select potential contributors, but not actually the competence of the contributor, and by extension their likelihood to mindfully, diligently, and accurately perform the task.

Given the findings in recent literature, we propose the following research question in order to evaluate the suitability of quality assurance measures in crowdsourcing:

***RQ**: What is the relationship between quality control and perceived response quality in microtasks?*

We hypothesize that the quality control method does not have a significant impact on contributor's response quality and that only in the complete absence of quality control we will find constantly underperforming contributors.

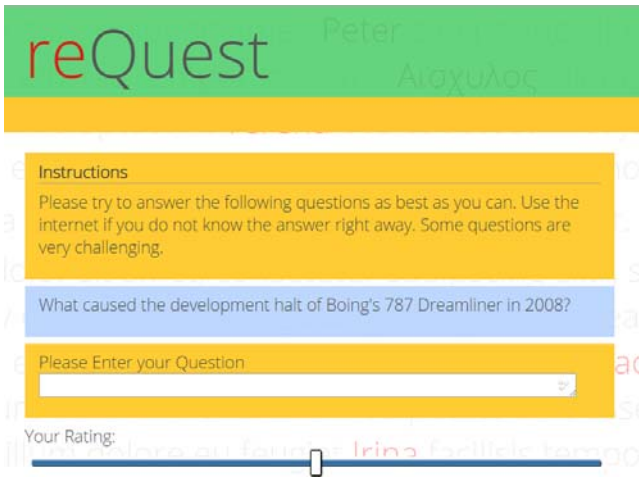

Figure 1: Crowdsourcing interface for the web-fragment annotation task. The interface is identical for all tasks. The rating slider (bottom) is only visible for our Raters when they judge the quality of a response.

Study Design

To investigate our question our study had a three (task complexity) by five (quality control methods) factorial and between groups design. The experiment investigates three tasks of varying complexity resulting in three levels of the *task* complexity factor. We hypothesize the order of tasks in terms of complexity to be as follows *semantic* similarity (least complex), *question* answering (more complex), and text *translation* (most complex).

We repeated each task five times with different methods of quality control. For the first level of the *control* factor (*none*) we did not perform any quality control. For the second level (*fake*) we announced very prominently in the task description that we use introductory quizzes to check the qualification of contributors, yet contributors did not get a test. The third level (*intro*) announces an introductory quiz and requires contributors to complete the quiz with 80% accuracy. In the fourth level (*auto*) we added a basic machine learning system to estimate the quality of a response and report this estimate to contributors. The system provides feedback on a three level scale (good, acceptable, unacceptable). Finally, in the fifth level (*wizard*) we replaced the ML-system by a human observer that decides on the response quality. The scale was identical to the one used by the ML-system.

We recruited all contributors via *crowdflower*. To control possibly confounding variables, provide feedback, and perform our own quality control we redirected contributors to our own webpage. After completing the task contributors get a code that they use to receive their payment through the *crowdflower* interface. The user interface was identical for all 15 (three by five) conditions. In all conditions contributors were shown three examples of correctly solved tasks and a description of the task. We also used the same interface to collect quality ratings from human judges. A screenshot of the interface as seen by the judges can be found in Figure 1.

To ensure a between groups design we used IP-tracking and browser fingerprinting to ensure that contributors do not contribute to more than one condition. To ensure contributor privacy only hashes of fingerprints and IP's were stored.

Automated Feedback

The automated feedback system applied in the level *auto* of the *control* factor requires some explanation. Experiments show that in some natural language tasks (Runge et al. 2012) the quality of a response can be estimated with a high accuracy by a combination of time needed to complete a single request and the numbers of characters typed. Although the values of both variables and their meaning differ from task to task, a machine learning classifier is able to learn the relationship between the two variables (features) and the response quality with minimal training data.

For our *auto* level, we classify responses into three different classes (good, acceptable, unacceptable) using a random forest classifier (Breiman 2001). Supervised classifiers need labeled training data. We classified 90 responses of each task by hand. We randomly selected responses and classified them into the three classes until there were 30 samples per class. We normalized the training data randomly, selecting exactly 30 samples per class.

For the experiment a random forest classifier was chosen, as tree-based classifiers are less sensitive to outliers and unbalanced sample sets (Cieslak & Chawla 2008). In the given tasks, it is likely that we encounter outliers such as a contributor opening a task and leaving her working place for a while. Classifiers such as support vector machines are more sensitive to such outliers. Our classifier generated 10 random trees using Gini impurity (Breiman 1996) as split criterion. We used the classifier that is part of pythons sklearn package (Pedregosa & Varoquaux 2011).

When the classifier estimates the response quality to be unacceptable we show a general warning that the response might need revision. If the response was acceptable, we did not show a message. For good responses, we showed a message stating that the response was of good quality. The messages appeared as a red text immediately after a contributor responded to a request.

Measurements

We consider two independent variables the quality *control* method and *task* complexity as well as one dependent variable perceived response *quality*. To measure perceived response quality we asked two human judges to rate each response on a scale from 0.0 (low quality) to 1.0 (high quality) in 10 increments. We calculated the average perceived response quality for each contributor as our measurement for *quality*. We consider contributors with an average perceived response quality below 0.6 as constantly underperforming. That means contributors with 40% unacceptable responses.

Judges used a web interface that was identical to the contributors interface. Judges saw the initial request and answer. Additionally judges had a slider to rate the response quality. The interface did not show the rating of our automated feedback system. We ensured that the process was blind. We randomly selected responses from all conditions and judges did not know the condition of a response. These judges were not involved in generating the training data for the automated feedback nor did they participate in the *wizard* conditions. We recruited the judges' offline from our lab.

We measure and report the agreement between judges using Krippendorff's Alpha (Krippendorff 1970). 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.

Furthermore, the three tasks are tested for instruction clearness and contributor satisfaction using the build in metrics

provided by *crowdflower*. Upon completion of a task, contributors can take a satisfaction survey. Contributors score the task on a 0-5 scale for overall satisfaction, instruction clearness, test question fairness, payment, and ease of job. Results of these quizzes are reported with each task.

Procedure

We collected all data for three independent tasks of varying complexity from the domain of natural language processing. The main interface for contributors is identical for all tasks. Figure 1 shows a screenshot of the user interface for the question-answering task. Table 1 shows the distribution of our contributors by level of quality control method and task complexity.

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

Table 1: Distribution of contributors over all fifteen conditions.

Word-based Semantic Similarity

Semantic similarity plays an important role for many natural language processing tasks, especially word sense disambiguation and information retrieval (Feng et al. 2008; Navigli 2009). Humans are better than algorithms at rating semantic similarity between two words (Batram et al. 2014). Involving paid online contributors can reduce costs, but the response quality is harder to predict. Constantly under-performing contributors are still an issue for such tasks (Krause & Porzel 2013). Different algorithmic approaches do exist (Strube & Ponzetto 2006; Yang & Powers 2005; Resnik 1995) but are not yet able to reproduce human level results (Radinsky et al. 2011). The task issued in this treatment is itself not very complex, only requiring a good command of the English language. To ensure this we restricted contributors origin to be in the US, UK, or Canada.

We further restricted the task using a standard dataset that was introduced by Finkelstein et al. (Finkelstein et al. 2001) that consists of 353 word pairs. In the experiment, we recruited 90 contributors and collected ~9,500 responses on the 353 word pairs.

Question Answering

Understanding natural language is still a challenging field for artificial systems (Krause 2014a). 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 (Ferrucci et al. 2010), unsolved challenges (Aras et al. 2010; Krause 2014b). Standard data sets for question answering seem too easy for human annotators with access to the internet. Therefore, we designed a set of 50 questions so that using the question as a search string does not 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 off-line from students and members of our research group. The response quality of a contributor is estimated by the semantic similarity between the contributor's response and our exemplary answers. We take 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) is 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 *crowdflower*. We collected 1,017 responses on average for each *control* level.

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 (Resnik et al. 2010) and iterative collaboration between monolingual users (Chang et al. 2010) are two examples. Other common approaches utilize mono- or bilingual speakers to proofread and correct Machine Translation results (Zaidan & Callison-burch 2011).

For our experiment, we use a popular Wikipedia article in German on the Brandenburg Gate. Native speakers of German 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

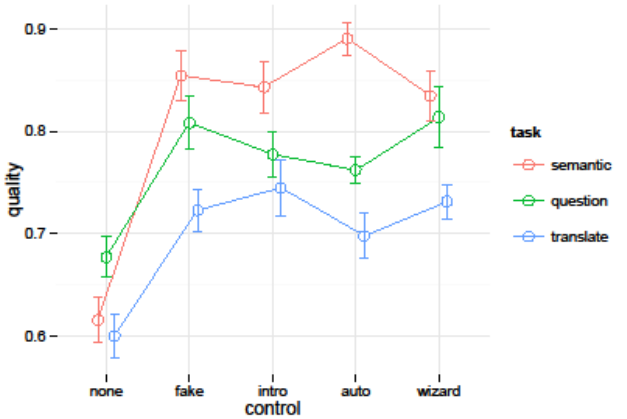

Figure 2: Perceived response quality for all fifteen conditions. Colors indicate the three different tasks. The lines are meant as visual aids. Error bars indicate standard error.

via *crowdflower*. 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.

We allowed each contributor to translate up to 100 sentences. We collected 2,119 translations for the Vietnamese set and 2,002 translations for the German set (total 4121) from 90 contributors (46 on average). We collected 825 sentences on average in each *control* condition.

Results

Before we analyze our data for main and interaction effects, we want to ensure that our presumption that the three different tasks have a distinct complexity is reasonable. We indeed found that the response quality is significantly lower for complex tasks. This indicates that the tasks do differ in their complexity. This is in line with the self-assessment of contributors through *crowdflowers* satisfaction survey. We found that *Ease Of Job* negatively correlates with our presumed complexity ranking. The correlation is significant with $p < 0.001$. Table 2 shows the results of the satisfaction survey.

	satisfaction	*clearness*	*fairness*	*payment*	*ease*
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

Table 2: Results of the self-assessment of our contributors on crowdflower. 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 crowdflower only offers aggregated data.

Additionally we ensure that our metric is reasonable. We use perceived quality as our measurement as this measure allows investigating quality over different tasks. Table 3 shows that our judges have a substantial agreement on quality throughout all tasks.

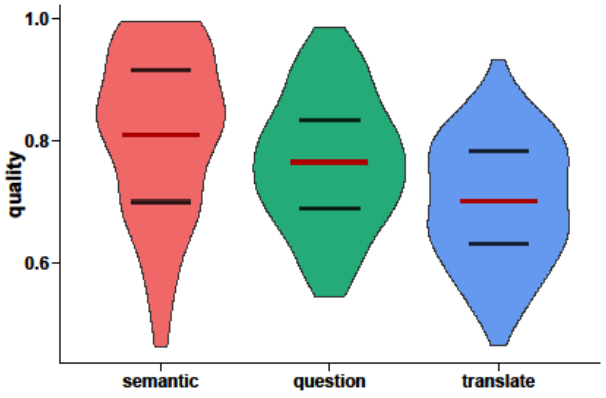

Figure 3: Task complexity affects response quality. The most complex task text translation (right) has a significantly lower average response quality than the more simplistic semantic similarity task (left) and the question answering task (middle). The figure shows a violin plot (Hintze & Nelson 1998)**. A violin plot combines a boxplot and a kernel density plot. Thick dark lines indicate 1st and 3rd quartiles the red lines population means.**

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

Table 3: Inter-rater agreement on perceived response quality. The results are homogenous for all three tasks and indicate a substantial agreement between our judges.

Before testing our results for significance, we ensured that our data is suitable for parametric tests. We used the Shapiro-Wilk test for normality (Royston 1982) for each condition and did not find significant differences from a normal distribution.

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) (Bartlett 1937). 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 analyze 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 4 shows the results of this test.

	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 4: ANOVA results of main and interaction effects. The first row shows the effect of the quality control method. The second row the effect of the task. The third row shoes the interaction effect between both factors.

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*. 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 5 shows the differences between levels of the *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	

Table 5: Results of Welch two sample t-tests with Holm correction comparing all levels of the quality control factor.

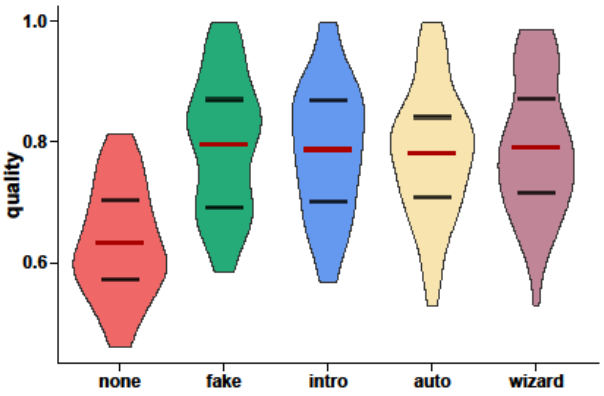

Figure 4: Quality control affects response quality yet only if there is no quality control at all. The differences in means between quality control methods are not significant.

Task Complexity Affects Response Quality

We also analyze effects for each level of the *task* complexity factor. We assume that the average response quality will deteriorate for tasks with higher complexity. As seen in Table 6 and Figure 3 this assumption holds for our experiment. Although this may seem obvious it illustrates that, the initial assumption on task complexity is accurate. The Pearson moment correlation is 1.0 with an associated $p < 0.001$.

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	***

Table 6: Results of Welch two sample t-tests with Holm correction. The first line compares level semantic to level question of the task complexity factor. Line two compares level semantic to translation and line three question to translation.

The Differences between Quality Control Methods are Insignificant

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 5. Other levels do not differ significantly. Table 7 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. We consider a contributor below a quality level of 0.6 constantly underperforming. We found that in all conditions with no quality control we had a substantial amount of contributors (N = 22) with an average response quality below 0.6. In all other conditions combined, we only 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.

	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

Table 7: Means and standard deviations for perceived quality. Rows contain the five different quality control methods while columns contain the different tasks of the experiment.

Machines can predict Response Quality almost as well as Humans

In the *auto* level of the quality *control* factor a 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 analyzed the correlation using Spearman's ρ. We found a substantial correlation between the predictions and the average perceived quality of our human judges $\rho(937020) = 0.71$, $p < 0.001$. The correlation between the two human judges in comparison is $\rho(463061) = 0.85$, $p<0.001$. In contrast, the human raters who replaced the ML-system in our wizard condition achieved a correlation of $\rho(705574) = 0.78$, $p<0.001$.

Conclusion

In this paper, we investigated the effect of different quality control methods on the response quality of contributors for tasks of varying complexity. We found as expected that our tasks differ in complexity and confirmed the hypothesized order to be as follows semantic similarity (least complex), question answering (more complex), text translation (most complex).

We found that constantly underperforming contributors (by our definition contributors with less than 40% acceptable responses) are almost not present in all conditions of our experiment with a quality control method in place. We however found a substantial amount of constantly underperforming contributors (almost 45%) in our control conditions (*none*) without a quality control method.

Only mentioning a required introductory test (without actually doing the test, the *fake* level of the *control* factor) was sufficient to achieve the same response quality as with other quality control methods. Even immediate human generated feedback was not able to raise response quality above the level of this faked introductory test. As hypothesized, the response quality does not differ across the different quality control methods. It only differs significantly between the *none* conditions (M = 0.63, SD = 0.03) and conditions with quality control (M = 0.79, SD = 0.05). This is an increase of more than 25% in response quality.

We therefore conclude that constantly underperforming contributors are aware of the fact that their contribution might fall short of required quality standards when taking a task. This also means that very basic quality control methods are sufficient to promote diligent work.

It is however debatable if our fake introductory test would keep these results over time. It is very likely that contributors realize that the tests are not conducted. However, we also demonstrated that extremely simple machine learning methods with task independent features as proposed by Krause et al. (Krause 2014b) can predict response quality on the fly. Such methods may provide quality control for tasks similar to the ones explored in this paper.

Limitations and Future Work

While a 3x5 factorial model is sizable, future work should cover more of the scope of quality control mechanisms to assure the transferability of these results. Furthermore, it has yet to be seen if tasks in other domains than natural language processing yield similar results.

As already noted we recognize that our minimal control mechanism (*fake*) without enforcement is not sustainable - contributors can and will quickly realize that no quality control has in fact been enforced. A sustainable and low cost mechanism to elevate the performance of diligent but underperforming contributors must be developed and tested to complete the scope of this research.

As shown in this work, after an even-basic controlling for response quality, underperformance per task drops considerably. A worthy area of future research is support systems for those who worked diligently but are still underperforming. This is both for the requestor's side (i.e., task description writing) as well as the contributor's side (i.e., educational materials).

Particularly worthwhile would be the investigation of monetary incentivization of contributors' education (see e.g., (Krause et al. 2016; Suzuki et al. 2016). Monetized education-based tasks could create the scenario that contributors are both learning to complete more and more complex tasks, while gaining skills and funding to be applied in their offline lives. An envisioned mechanism for this could be Massively Open Online Courses, where contributors register for the course to learn increasingly complex skills, and are financially rewarded with successful task mastery. Realized in its full depth and scope, this progressive step would contribute to the comprehensive enhancement of both crowdwork from a quality perspective and the overall, real life skillset of the contributors.

References

Aras, H. et al., 2010. Webpardy: Harvesting QA by HC. In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, New York, USA: ACM Press, pp. 49–53.

Bartlett, M., 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series*, A(160), pp.268–282.

Batram, N., Krause, M. & Dehaye, P., 2014. Comparing Human and Algorithm Performance on Estimating Word-based Semantic Similarity. In *SoHuman'14 Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation*. Barcelona, Spain: Springer Berlin Heidelberg, pp. 131–139.

Breiman, L., 2001. Random Forests. *Machine learning*, 45(1), pp.5–32.

Breiman, L., 1996. Technical note: Some properties of splitting criteria. *Machine Learning*, 24(1), pp.41–47.

Chang, H., Bederson, B.B. & Resnik, P., 2010. MonoTrans2: An Asynchronous Human Computation System to Support Monolingual Translation. *hcil.cs.umd.edu*, pp.2–5.

Chen, J.J. et al., 2011. Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Interfaces*, 5(3).

Cieslak, D. & Chawla, N., 2008. Learning decision trees for unbalanced data. In *European Conference on Machine Learning*. Antwerp, Belgium: Springer.

Dow, S. et al., 2012. Shepherding the crowd yields better work. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, p.1013.

Dukat, C. & Caton, S., 2013. Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities. In *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*. Karlsruhe, Germany, pp. 536–540.

Feng, J., Zhou, Y. & Martin, T., 2008. Sentence Similarity based on Relevance. In *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '08)*. pp. 832–839.

Ferrucci, D. et al., 2010. Building Watson: An overview of the DeepQA project. *AI magazine*, 31(3), pp.59–79.

Finkelstein, L. et al., 2001. Placing Search in Context : The Concept Revisited. , pp.406–414.

Geiger, D. et al., 2011. Managing the Crowd : Towards a Taxonomy of Crowdsourcing Processes. In *Proceedings of the 17th Americas Conference on Information Systems*. pp. 1–11.

Hintze, J.L. & Nelson, R.D., 1998. Violin plots: A box plot-

density trace synergism. *American Statistician*, 52(2), pp.181–184.

Ipeirotis, P.G., Provost, F. & Wang, J., 2010. Quality management on amazon mechanical turk. In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*. Washington, DC, USA: ACM Press, pp. 0–3.

Kern, R., Thies, H. & Satzger, G., 2010. Statistical Quality Control for Human-based Electronic Services. In *Proceedings of Service-oriented computing: ICSOC 2010*. San Francisco, CA, USA: Springer Berlin Heidelberg, pp. 1–17.

Kittur, A. et al., 2013. The Future of Crowd Work. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. p. 1301.

Krause, M. et al., 2016. Connecting Online Work and Online Education at Scale. In *CHI Extended Abstracts on Human Factors in Computing Systems*. pp. 3536–3541.

Krause, M., 2014a. Designing Systems with Homo Ludens in the Loop. In P. Michelucci & K. Greene, eds. *Handbook of Human Computation*. New York, NY, USA: Springer New York, pp. 393–409.

Krause, M., 2014b. *Homo Ludens in the Loop: Playful Human Computation Systems*, Hamburg, Germany: tredition GmbH, Hamburg.

Krause, M. & Porzel, R., 2013. It is about time. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. New York, New York, USA: ACM Press, p. 163.

Krippendorff, K., 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(61), pp.61–70.

Navigli, R., 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41, p.10.

Oleson, D. et al., 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *HComp'11 Proceedings of the AAAI Workshop on Human Computation*, pp.43–48.

Pedregosa, F. & Varoquaux, G., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Radinsky, K. et al., 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International World Wide Web Conference WWW'11*. Hyderabad, India: ACM Press, pp. 337–346.

Resnik, P., 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. p. 6.

Resnik, P. et al., 2010. Using Monolingual Human Computation to Improve Language Translation via Targeted Paraphrase. In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*. Washington, DC, USA: ACM Press.

Royston, J.P., 1982. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal Of The Royal Statistical Society Series C-Applied Statistics*, 31, pp.115–124.

Runge, N. et al., 2012. Predicting Crowd-Based Translation Quality with Language-Independent Feature Vectors. *Workshops at the Twenty- …*, pp.114–115. Available at: http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/viewPaper/5237.

Sarasua, C. & Thimm, M., 2013. Microtask available, send us your CV! In *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*. pp. 521–524.

Sheng, V.S., Provost, F. & Ipeirotis, P.G., 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 614.

Stolee, K.T. & Elbaum, S., 2010. Exploring the use of crowdsourcing to support empirical studies in software engineering. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, p. 35.

Strube, M. & Ponzetto, S.P., 2006. WikiRelate ! Computing Semantic Relatedness Using Wikipedia. In *AAAI*. pp. 1419–1424.

Suzuki, R. et al., 2016. Atelier : Repurposing Expert Crowdsourcing Tasks as Micro-internships. In *Chi 2016*.

Wang, J., Ipeirotis, P. & Provost, F., 2013. Quality-Based Pricing for Crowdsourced Workers. , pp.1–46.

Yang, D. & Powers, D.M.W., 2005. Measuring Semantic Similarity in the Taxonomy of WordNet. *Reproduction*, pp.315–322.

Zaidan, O.F. & Callison-burch, C., 2011. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of ACL 2011*. pp. 1220–1229.