An Empirical Study on Short- and Long-Term Effects of Self-Correction in Crowdsourced Microtasks

Masaki Kobayashi
University of Tsukuba
makky@klis.tsukuba.ac.jp

Hiromi Morita
University of Tsukuba
morita@slis.tsukuba.ac.jp

Masaki Matsubara
University of Tsukuba
masaki@slis.tsukuba.ac.jp

Nobuyuki Shimizu
Yahoo! Japan
nobushim@yahoo-corp.jp

Atsuyuki Morishima
University of Tsukuba
mori@slis.tsukuba.ac.jp

Abstract

Self-correction for crowdsourced tasks is a two-stage setting that allows a crowd worker to review the task results of other workers; the worker is then given a chance to update his/her results according to the review. Self-correction was proposed as an approach complementary to statistical algorithms in which workers independently perform the same task. It can provide higher-quality results with few additional costs. However, thus far, the effects have only been demonstrated in simulations, and empirical evaluations are needed. In addition, as self-correction gives feedback to workers, an interesting question arises: whether perceptual learning is observed in self-correction tasks. This paper reports our experimental results on self-corrections with a real-world crowdsourcing service. The empirical results show the following: (1) Self-correction is effective for making workers reconsider their judgments. (2) Self-correction is more effective if workers are shown task results produced by higher-quality workers during the second stage. (3) Perceptual learning effect is observed in some cases. Self-correction can give feedback that shows workers how to provide high-quality answers in future tasks. The findings imply that we can construct a positive loop to improve the quality of workers effectively. We also analyze in which cases perceptual learning can be observed with self-correction in crowdsourced microtasks.

1 Introduction

Ensuring the quality of obtained data is one of the primary problems in crowdsourcing; numerous studies have addressed the problem of improving the quality of task result data. In particular, for the categorization/labeling task, which is considered to account for a large portion of microtasks in a crowdsourcing service such as Amazon Mechanical Turk, three approaches are commonly used.

The first is to choose good workers. For example, with Amazon Mechanical Turk, most requesters attempt to recruit workers with high approval ratings or category masters selected by the platform. The second approach is to assign the same task to multiple workers and aggregate the results, which allows the final results to be computed using various aggregation methods (e.g., majority voting).

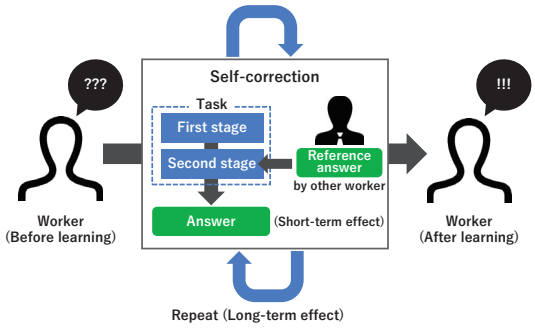

Figure 1: With self-correction, after workers answer the task question (first stage), they review other workers answers (second stage) and reconsider their answers in light of this review. In this study, we evaluate self-correction by workers, focusing on both short- and long-term effects.

The third approach is to derive better results from individual workers. Shah and Zhou proposed a two-stage setting for crowdsourced tasks, named self-correction, which shows other workers' task results to each worker after the results are submitted, allowing the worker to update his/her results (Shah and Zhou 2016). Self-correction can be incorporated into crowdsourcing tasks performed on commercial crowdsourcing services as an external task.

Shah and Zhou argued that self-correction is effective, particularly when workers perform poorly in the first stage. The point here is that the workers notice the mistakes they made in the first stage, and subsequently correct them in the second stage. Self-correction provides workers with an opportunity to notice their mistakes. Importantly, self-correction is complementary to result-aggregation methods in which multiple workers independently perform the same task, and the results are aggregated. Self-correction thus derives better quality with few additional costs.

However, the effectiveness of self-correction has only been shown in simulations and is not supported by evidence from real-world experiments. Therefore, whether it is effective in real-world settings is an interesting question that deserves attention.

Another interesting question is whether we can observe involuntary perceptual learning effects in self-correction microtasks. If workers perform a sequence of self-correction tasks, it is unclear whether feedback helps them improve the results of future tasks. If feedback is helpful, this implies that there are cases where we can increase the quality of workers without explicit training phases with known so-called gold-standard data.

It is known that the ability to perform a perceptual task is improved by repetition, i.e., perceptual learning (Gibson 1969). Perceptual learning occurs even involuntarily (Gibson and Gibson 1955), and some studies have reported perceptual learning in visual categorization tasks (Mettler and Kellman 2014). Therefore, we expect that repeating self-correction tasks will improve workers' ability to perform visual categorization tasks.

This paper reports the results of experiments designed to examine the short- and long-term effects of self-correction in a real crowdsourcing setting (Figure 1). Our key findings are as follows:

1. The short-term effects of self-correction on data quality are observed in a real-world setting. Here, the short-term effects are data quality improvements in the same task. That is, workers notice mistakes they made in the first stage.
2. We found that the quality of data shown to workers in the second stage is important. Indeed, self-correction is more effective if workers are shown the task results produced by higher-quality workers in the second stage.
3. The long-term effects of self-correction on data quality are also observed. Here, the long-term effects were perceptual learning by workers, namely quality improvements in a successive sequence of similar but different tasks. This result suggests that self-correction teaches workers how to give high-quality answers during similar subsequent tasks.

In addition, we analyzed in which cases perceptual learning can be observed with self-correction in crowdsourced microtasks.

All our findings were obtained using a simple monetary incentive. Each worker was paid according to the number of the tasks performed. In (Shah and Zhou 2016), the authors designed a monetary incentive that was proven to be theoretically optimal. Interestingly, our results showed that self-correction is effective for improving the quality of both task results and workers, even with this simple monetary incentive. Introducing a more sophisticated incentive is one of our future research topics. However, such an incentive may limit the crowdsourcing services we can use with self-correction.

2 Related work

Improving the quality of data and workers is the focus area in crowdsourcing, and numerous studies have addressed these issues (Daniel et al. 2018).

One major approach is to assign the same task to multiple workers and aggregate the results to obtain the final results. With this approach, high-quality results are expected. Majority voting is the most straightforward strategy for aggregating the results; however, more sophisticated aggregation strategies have been proposed that depend on a variety of factors, such as the quality of the workers, agreement ratios, and clustering results (Hung et al. 2013) (Jagabathula, Subramanian, and Venkataraman 2014) (Aroyo and Welty 2013). Other approaches include those addressing better task design (Doroudi et al. 2016) and incentive structures (Kinnaird et al. 2013) (Hsieh and Kocielnik 2016). Note that self-correction can be combined with any of them.

There are a variety of studies focusing on worker feedback, and from these, we know that feedback can improve the quality of task results. Revolt (Chang, Amershi, and Kamar 2017) and Microtalk (Drapeau et al. 2016) give workers opportunities to change their answers after seeing justifications of other workers answers. Shepherd (Dow et al. 2012) allows both self-assessment and external assessments of various forms. Self-correction proposed by (Shah and Zhou 2016) offers a simple form of feedback, which is someone else's answer to the same question. However, exactly how this feedback works from a scientific perspective remains to be studied.

It is known that assessments by workers are biased. Gadiraju et al. showed that crowd workers often lack awareness about their actual level of competence (Gadiraju et al. 2017). Incorporating such bias into a self-correction framework would be an interesting challenge for future work.

If we focus on improving worker quality for better task results, a typical approach will be to train workers before asking them to perform tasks. It is known that workers perform better if they are asked to perform tasks for training purposes before they perform normal tasks (Ashikawa, Kawamura, and Ohsuga 2015). Such an approach requires us to prepare training tasks for the workers, and to know the answers in advance to teach them to the workers. Suzuki et al. proposed something call micro-internships that connect intern workers to experts through crowdsourcing for supporting them in obtaining the required skills (Suzuki et al. 2016).

It is expected that perceptual learning occurs in microtasks with feedbacks. Abad et al. showed that rule-based feedback given to workers who provided incorrect answers is effective for training workers (Abad, Nabi, and Moschitti 2017). Our question is whether this happens even with a simple form of feedback. Below, we experimentally demonstrate that there are cases where perceptual learning is observed, and workers are trained to produce higher-quality work while they perform ordinary tasks in a self-correction framework. An important key to perceptual learning is how many times each worker performs the same tasks. Law et al. (Law et al. 2016) discussed incentive structures for keeping workers engaged in the same tasks for a long time. It would be interesting to introduce such a mechanism into our framework.

Below, we show that, with self-correction, improvements in the quality of task results (i.e., short-term effects) are more evident when task results produced by high-quality workers are shown in the second assignment. However, self-correction does not specify how to find high-quality work-

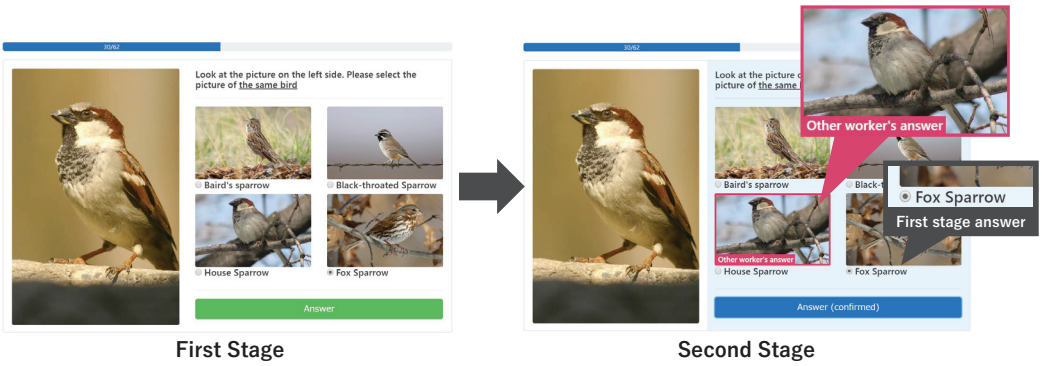

Figure 2: Self-correction tasks: In the first stage, a worker answers a question. In the second stage, the worker corrects his/her answer by reviewing other workers' answers.

ers when we choose the answers for the second assignment. The key here is an index that measures the quality of workers. There are many approaches to measuring worker quality (Hung et al. 2015) (Haas et al. 2015) (Gadiraju et al. 2015). The simplest approach is to insert special tasks into the sequence of normal tasks for measuring worker quality. It is shown in (Jung and Lease 2015) that to ensure accurate measurements, tasks for measuring worker quality should be used not only in the first several task assignments but also regularly in the continual sequence of task assignments. Moreover, numerous research efforts have focused on techniques for measuring worker quality without the use of so-called gold standard data (Joglekar, Garcia-Molina, and Parameswaran 2013) (Das Sarma, Parameswaran, and Widom 2016).

3 Self-correction

In this section, we describe the self-correction proposed in (Shah and Zhou 2016). The important features of their proposed method are as follows:

In typical crowdsourcing services, workers do not have the opportunity to find their errors. If we provide workers the opportunity to notice their errors, they will be able to correct their answers (if they are not spam workers). Self-correction is a task designed to improve the quality of output from crowd workers. With self-correction, a worker answers the same question twice (Figure 2). During the first stage, the worker offers an initial answer. Then, during the second stage, the worker can revise this answer after reviewing other workers' answers.

In the self-correction process, some workers may not work seriously in the first stage because they must answer without considering answers obtained from others during the second stage. Thus, Shah et al. proposed an incentive algorithm for self-correction settings. To prevent worker carelessness, workers receive rewards when they answer questions correctly during the first stage.

A self-correction simulation was conducted to clarify its usefulness. In the simulation, a standard task was compared to a task in which self-correction was applied. The results of the experiment showed that more accurate output could be obtained from a worker performing a task that included self-correction. According to the authors, the error rate of the machine-learning algorithm could be reduced by using a dataset obtained via self-correction for learning.

4 Experiment 1

4.1 Research questions

We conducted Experiment 1 to investigate the following research questions: (1) Do the self-correction tasks improve the quality of answers in real-world crowdsourcing settings? (Short-term effect) (2) Do we need other worker's answer as reference answer in self-correction tasks? (Trusted vs. Self) (3) Does repeating self-correction tasks induce an involuntary learning effect in workers? (Long-term effect)

4.2 Participants

One-hundred ninety-six workers participated in the experiment through Yahoo! Crowdsourcing [1].

The task instruction was in Japanese because all the workers who participated were Japanese / understood Japanese. Each Yahoo! Crowdsourcing task consisted of many classification tasks such as those shown in Table 2. The classification tasks were generated by using Crowd4U [2].

To investigate the effectiveness of reference answers, the workers were divided into two groups (Table 1). Half of them were assigned to a group that engaged in self-correction with a reference answer (hereafter, "trusted"). The other half was assigned to a group engaging in self-correction without a reference answer (hereafter, "self"). The workers were to receive a reward of about $1 when they completed all the tasks.

Note that each self-correction task consisted of two-stages for both trusted and self groups. In the second stage, the workers can reconsider their own answer of the first stage either by reviewing other workers' answer as a reference in the trusted group or without the reference in the self group.

[1] https://crowdsourcing.yahoo.co.jp
[2] https://crowd4u.org

Table 1: Experiment settings

	Experiment 1	Experiment 2
Group conditions	Trusted vs. Self	Correct vs. Random
Learning phase	28 tasks	52 tasks + 2 gold standard questions
Image dataset	Caltech-UCSD Birds 200	Pictures from wikiart.org
Filter	Under 25% in mid or post phase	Gold standard questions

Table 2: Procedure

	Phase	Task type	Number of tasks
1	Pre-test	Test	12
2	Learn 1	Self-correction	(Follow Table 1)
3	Mid-test	Test	12
4	Learn 2	Self-correction	(Follow Table 1)
5	Post-test	Test	12

4.3 Procedure

Workers were asked to perform three phases of test tasks and two phases of self-correction tasks, as shown in Table 2. Pre-test, mid-test, and post-test phases were designed for obtaining worker ability assessments. In these phases, workers were asked to perform 12 test tasks. By comparing the accuracy rate of the results from the pre-test, mid-test, and post-test phases, we could clarify the involuntary learning effects of self-correction tasks. In the two self-correction phases, workers were asked to perform 28 self-correction tasks. By comparing the accuracy rate of the results from the first and second self-correction stages, we could verify the quality improvement effects of self-correction.

4.4 Tasks

Test phase and the first stage of the self-correction phase In the test phase and first stage of the self-correction phase, we used a four-class classification task (Left side in Figure 2). The classification task involved answering a question by selecting a particular image. We displayed an image of an unidentified bird on the left side of a screen. On the right side, we presented four photos of birds with names underneath. From these choices, workers were asked to identify the bird in the image on the left.

To avoid the ceiling effect, we kept the tasks difficult. That is, all birds were chosen from similar classes. Among all tasks, question images were used only once. We collected four bird images from Caltech-UCSD Birds 200 (Welinder et al. 2010). We carefully selected four types of birds with similar characteristics.

Second stage of self-correction phase In the second stage of the self-correction phase, we presented the question image and worker's choice from the first stage again. Workers could change their choice. We highlighted other workers' answers as reference answers in the experiment involving self-correction with reference answers (Right side in Figure 2).

4.5 Reference answers

In Experiment 1, the reference answers were obtained from the top 20% of the highest-scoring workers in the "self group". Each worker in the "trusted group" was randomly paired with a worker from the set of top 20% workers.

4.6 Filter

To exclude the data of underperforming workers who may have been making random choices because of fatigue or satisfice, we disregarded the data of workers whose accuracy rate did not exceed 25 percent in at least one of the mid- or post-phases. As a result, 86 workers in the "trusted group" and 84 workers in the "self group" remained for analyses (Figure 3).

4.7 Results

Table 3 shows the accuracy in the pre-test phase. There is no significant difference between the average of accuracy of self and trusted groups.

Short-term effect Figure 3 shows the accuracy rate at two stages of self-correction of each group.

We conducted a two-way ANOVA with the accuracy rate as a dependent variable, the stage as a within-worker factor, and the presentation of reference answers as a between-worker factor. As a result, there were significant effects from the stage ($F(1, 168) = 39.321$, $p < .001$), and from the type of reference answer ($F(1, 168) = 10.454$, $p < .001$) and their interaction ($F(1, 168) = 48.290$, $p < .001$).

We proceeded with post-hoc analyses because the interaction was significant. There was a simple main effect from the stage for workers in the "trusted group" ($F(1, 168) = 88.42$, $p < .001$), but no main effect from the stage for workers in the "self group" ($F(1, 168) = .23$, n.s.). There was no simple main effect from the type of reference answers for the first stage ($F(1, 168) = 0.18$, n.s.), but there was a simple main effect from the type of reference answers for the second stage ($F(1, 168) = 31.82$, $p < .001$).

Long-term effect Figure 4 shows the accuracy rate of the test phase of each group.

We conducted a two-way ANOVA with the accuracy rate as a dependent variable, the test phase as a within-worker factor, and the type of reference answer as a between-worker factor. The ANOVA revealed a significant effect from the test phase ($F(2, 336) = 8.731$, $p < .001$), but no effect from the type of reference answer ($F(1, 168) = 0.635$, n.s.).

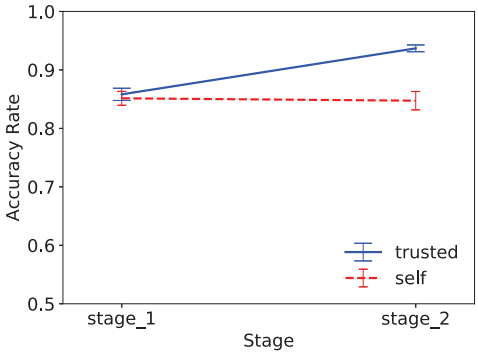

Figure 3: [Experiment 1] Accuracy rate with the first and second stages of each learning phase.

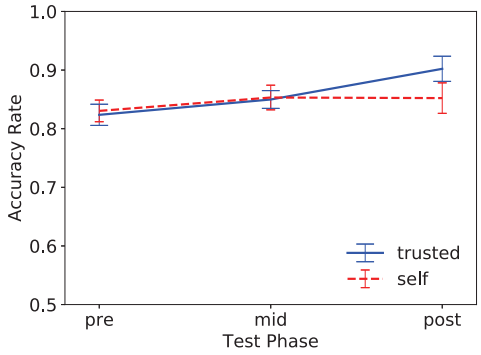

Figure 4: [Experiment 1] Accuracy rate with each test phase.

Table 3: Pre-test phase accuracy in Experiment 1

Condition	Filter	N	Median	Mean	Std
self	None	98	0.833	0.826	0.147
	Under 25%	84	0.833	0.83	0.136
trusted	None	98	0.833	0.816	0.131
	Under 25%	86	0.833	0.824	0.134

Because there was a significant interaction between the test phase and reference answers ($F(2, 336) = 3.50$, $p < .05$), we conducted post-hoc analyses that showed a simple main effect from the test phase for workers in the "trusted group" ($F(2, 336) = 11.19$, $p < .001$), but no simple main effect from the test phase for workers in the "self group" ($F(2, 336) = 1.14$, n.s.). A multiple-comparison using a Bonferroni correction with the "trusted answer rates for the "trusted group" revealed a marginal difference between the pre-test and mid-test rates ($p < .10$), and significant differences between the mid-test and post-test rates ($p < .005$), and between the pre-test and post-test rates ($p < .001$).

There was no simple main effect from the reference answers for the pre-test and mid-test ($F(1, 168) = 0.10$, n.s.; $F(1, 168) = 0.03$, n.s.) , but there was a simple main effect from the post-test ($F(1, 168) = 4.48$, $p < .05$).

4.8 Discussion

Short-term effect The accuracy rate of the workers in the trusted group increased in the second stage, while that of the workers in the self group did not. This means that referring to other workers' answers is an important factor in increasing the accuracy rate of the second stage answers because there was no difference between the accuracy rate of the first stage answers for both worker groups. Thus, it was proven that the self-correction strategy of referring to an appropriate reference answer increases the quality of worker output.

Long-term effect There was no difference in worker ability between groups at the beginning because there was no difference in performance between groups in the pre-test. However, there appeared a difference in performance between groups in the post-test, which means that workers developed an ability to give high-quality answers by repeating self-correction tasks with appropriate reference answers.

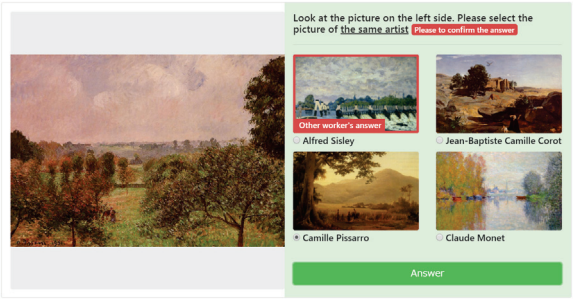

Figure 5: The second stage of self-correction task for Experiment 2. The workers were asked to classify paintings by artists.

5 Experiment 2

5.1 Research questions

We conducted Experiment 2 to investigate the following research questions: (1) Does the self-correction strategy work well with more difficult tasks? (2) How involuntary learning effect differs if the worker performs more tasks?

5.2 Participants

One-hundred ninety-one workers participated in an experiment through Yahoo! Crowdsourcing. In addition, actual tasks were assigned from Crowd4U as the external task. The workers were divided into two groups (Table 1). Half of them were assigned to a group that engaged in self-correction with a correct answer (hereafter, "correct"). The other half was assigned to a group engaging in self-correction with a random answer (hereafter, "random"). The workers were to receive a reward of about $1 when they completed all the tasks.

5.3 Procedure

Because the task was very difficult in Experiment 2, we assigned workers more self-correction tasks to provide them

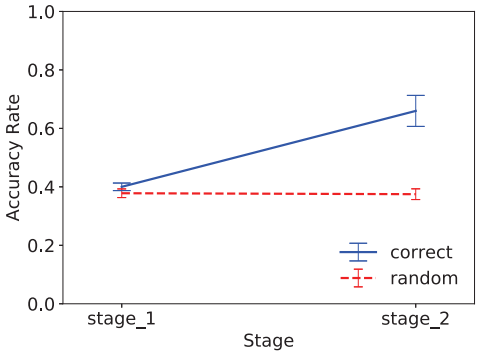

Figure 6: [Experiment 2] Accuracy rate with the first and second stages of each learning phase.

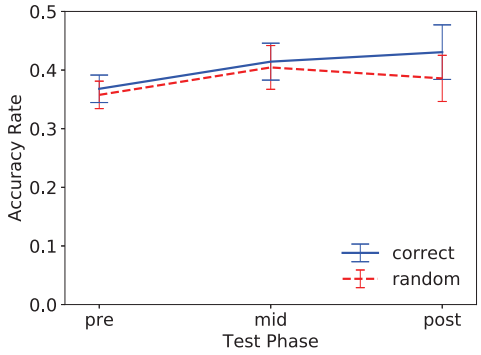

Figure 7: [Experiment 2] Accuracy rate with each test phase.

Table 4: Pre-phase accuracy in Experiment 2

Condition	Filter	N	Median	Mean	Std
random	None	105	0.333	0.354	0.152
	Gold	89	0.333	0.357	0.153
correct	None	86	0.333	0.355	0.15
	Gold	72	0.333	0.368	0.153

more learning opportunities. One learning phase contained 52 self-correction tasks (Table 2).

5.4 Tasks

The same task as in Experiment 1 was used except that workers were asked to classify paintings by four famous artists instead of bird images (Figure 5). The painting images were gathered from wikiart.org [3].

5.5 Reference answers

In Experiment 2, correct answers were used as reference answers in the second stage. We did not use trusted worker's answers as the reference answer because the task was so difficult that very few workers were expected to be able to give fully trusted answers.

5.6 Filter

In Experiment 1, some of the workers were not answering seriously at some point. In Experiment 2, we added a gold-standard question to the task list to find and omit such answers from the analysis. In the gold-standard question, one of the four exemplar paintings was presented to be classified. We analyzed only those workers who correctly answered the gold-standard questions.

5.7 Results

Table 4 shows the accuracy in the pre-test phase. Because the task was more difficult, the accuracy rates are lower than the accuracy rates obtained in Experiment 1.

[3]https://www.wikiart.org/

Short-term effect Figure 6 shows the accuracy rate of the two stages of self-correction for each group.

We conducted a two-way ANOVA with the accuracy rate as a dependent variable, the stage as a within-worker factor, and the quality of reference answers as a between-worker factor. As a result, there were significant effects from the stage ($F(1, 159) = 100.855$, $p < .001$) , the quality of reference answers ($F(1, 159) = 53.809$, $p < .01$) , and their interaction ($F(1, 159) = 106.02$, $p < .001$) .

We proceeded with post-hoc analyses because the interaction was significant. There was a simple main effect from the stage for workers in the correct group ($F(1, 159) = 187.09$, $p < .001$) , but no simple main effect from the stage for workers in the random group ($F(1, 159) = .04$, n.s.) . There was no simple main effect from the quality of reference answers for the first stage ($F(1, 159) = 1.38$, n.s.) , but there was a simple main effect from the quality of reference answers for the second stage ($F(1, 159) = 95.44$, $p < .001$).

Long-term effect Figure 7 shows the accuracy rate of the test phase of each group.

We conducted a two-way ANOVA with the accuracy rate as a dependent variable, the test phase as a within-worker factor, and the quality of reference answers as a between-worker factor. The ANOVA revealed a significant effect from the test phase ($F(2, 318) = 5.201$, $p < .05$), but no effect from the quality of reference answers ($F(1, 159) = 0.97$, n.s.) and their interaction ($F(2, 318) = 0.74$, n.s.).

Though there was no interaction, we conducted multiple-comparison for each group, since we were interested in the difference in the size of improvement between two groups. A multiple-comparison using a Bonferroni correction with accuracy rates for the correct-reference group revealed a significant difference between the pre-test and post-test ($p < .05$), and marginally significant difference between the pre-test and mid-test ($p = 0.052$) and no significant difference between the mid-test and post-test ($p = 0.51$). In contrast, a multiple-comparison with accuracy rates for the random-reference group showed no significant difference between the pre-test and post-test ($p = 0.21$) and between the mid-test and post-test ($p = 0.40$), but revealed a significant difference between the pre-test and mid-test ($p < .05$).

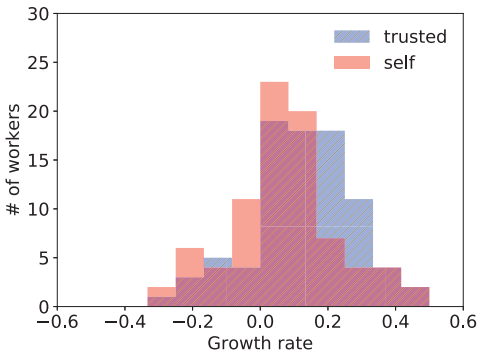

Figure 8: Distribution of workers growth rate with each condition in Experiment 1.

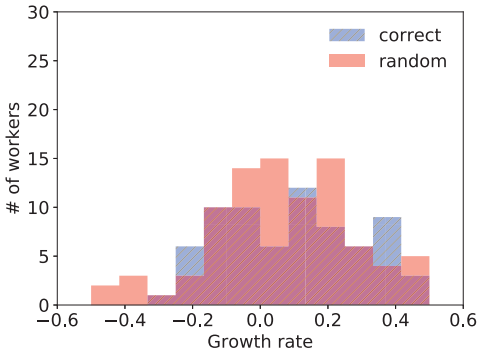

Figure 9: Distribution of workers growth rate with each condition in Experiment 2.

5.8 Discussion

Short-term effect The results of Experiment 2 supported a claim that has been shown in Experiment 1. In self-correction, presenting the correct reference at the second stage did improve the results, but presenting a random answer did not. However, it should be noted that some workers might have always adopted the reference answer in the second stage and obtained good results in the correct-reference condition. They will affect the growth rate, which will be described later.

Long-term effect The accuracy rate of the correct-reference group slightly improved between the pre-test and post-test. In contrast, the results of the random-reference group did not differ between pre-test and post-test. It increased between pre- and mid-tests, then decreased between mid- and post-tests. The pattern of results overall supports the research question about the long-term effect of involuntary learning, though the effect is quite trivial.

6 Discussion

6.1 Effect of self-correction

The effect of self-correction (Shah and Zhou 2016) involves two stages for crowdsourcing, in which a worker first answers a question, and then is allowed to change it in the second stage after reviewing a reference answer. According to numerical experiments, this setting is effective in overcoming various types of misjudgement commonly observed in crowdsourcing, provided that an elaborate incentive mechanism is used. In our experiment, real crowd workers performed tasks with self-correction. The experimental results revealed that the task results obtained from self-correction are more accurate. Thus, for more accurate answers, it is important to provide other workers' answers as reference answers during the second stage of self-correction. Presenting more accurate reference answers (or correct answers in the experiment) ensures the effect of self-correction.

This differs from original self-correction in that a fixed amount of fee is paid to the workers. However, even with our simple incentive mechanism, we did not observe malicious users. In fact, on average, in the second stage, approximately 10% workers changed their answer in Experiment 1 and approximately 25% changed their answer in Experiment 2, which shows that self-correction is effective even in such a simple setting.

6.2 Long-term effect of self-correction

By comparing worker performance in the pre-test and post-test phases, we found small but significant long-term involuntary learning effect. That is, the workers improved their ability by repeating self-correction tasks with a reference. The quality of task result is one of the most important topics in crowdsourcing. Retaining learned workers and assigning additional tasks to them is one way to support the problem. However, training requires workers to attend training session for a certain period before performing tasks and getting rewards. Therefore, the long-term learning effect of repeated self-correction with references is valuable because novice workers can advance into expert worker category while completing tasks and receiving rewards. The development will require a certain amount of time or a certain number of repetitions because the difference in accuracy rate for workers in the "trusted" group and those in the "self" group appeared only in the post-test phase in the bird categorization task. For the painting categorization task, we observed a difference in accuracy rate for the workers in "correct" group between the pre-test and post-test phases. Thus, although there was no difference in performance between the two groups in the post-test phase, we expect the difference to appear after performing more self-correction tasks.

Figure 8 shows the relationship between the worker growth rate and the number of the workers in Experiment 1. The x-axis shows the growth rate calculated by subtracting the correct rate of the pre-test from that of the post-test. The y-axis shows the number of workers. The distribution of the growth rate in the trusted condition shows that certain percentage of workers improved their results by 0% – 30% and few workers had a negative value. In contrast, the

distribution for the self-condition shows that there are many workers with growth rate of -10% – 20%.

Similarly, Figure 9 shows the relationship between the growth rate and the number of workers in Experiment 2. It seems that the distribution of growth rate does not differ between correct and random conditions. There are more workers with growth rate of 30% – 50% for the correct condition, but more workers with growth rate of -30% – 10%, too.

To find a good worker, it is important to find a worker who is already good. Many studies have focused on this problem. However, there are workers who have the potential for future excellence. We believe it is important to spot such potential excellence at an early stage (for example, during the pre-test and the learn 1 phases) by noting the answer patterns and reaction time.

6.3 Incentives for workers

For self-correction, a reward algorithm for workers who behave seriously has been proposed. However, in our experiments, we did not use this algorithm. We paid a fixed remuneration to the workers in our experiments. We proved that this setting works well in real crowdsourcing platforms without such elaborate incentive mechanisms. Nevertheless, if dynamic rewards are set within the crowdsourcing platform, workers may work more seriously when such a strategy is combined with existing reward algorithms.

Tasks that continue over long periods can render some workers fatigued or bored. They will begin to disregard the reference answer, give inadequate consideration to their second answers, and miss the chance to develop. Therefore, mechanisms are required that can assign tasks to the same worker for long periods. Such mechanisms should encourage motivated workers to continue tasks while guiding fatigued or bored workers to drop out.

6.4 Choosing workers for reference answers

In our experiment, the accuracy rate of the task was used as an indicator for selecting a worker for the reference answer. We used an image dataset with the correct label. However, finding the correct answer in a real crowdsourcing setting is often challenging. The simplest method of choosing a worker is to evaluate the performance of workers using simple tasks that are unrelated to the actual tasks.

Moreover, there are several methods to measure worker quality without the gold-standard data. We can use such methods to choose new workers whose answers are used in the second stage. The relevant research section in this paper refers to studies measuring workers' abilities. Identifying the best or better combinations for finding better workers is an interesting research topic. Although addressing this issue is beyond of the scope of this paper, we expect that the answer will depend on the nature of the tasks.

6.5 Deployment

There is more than one way to deploy self-correction into a set T of tasks to be submitted to typical commercial crowdsourcing platforms. A simple framework can be configured as follows: First, ask workers to perform a small set of test tasks (with gold-standard data) to measure the quality of workers. Ask the top $X\%$ of workers (denoted by E) to perform a subset of T without the second stage to obtain answers to be used in the second stage for others. Ask other workers to perform the same set of tasks in the second stage. Obtain answers for the last $n\%$ of tasks to choose workers whose answers were similar to those of E; they may be included in E for the next batch of tasks.

If the crowdsourcing platform allows it, we can consider combining dynamic reward algorithms and techniques to find good workers.

7 Conclusion

We reported our experimental results on self-corrections with a real-world crowdsourcing service. The results empirically showed the following:

- Self-correction is effective for making workers reconsider their judgments.
- Self-correction is more effective when workers are shown task results produced by higher-quality workers in the second stage.
- Perceptual learning is observed in some cases. Self-correction can provide feedback that helps workers provide high-quality answers for future tasks.

The findings imply that we can construct a positive loop to effectively improve the quality of workers. We also analyzed in which cases involuntary perceptual learning effects with self-correction were observed in crowdsourced microtasks. This analysis revealed that even difficult workers would improve in self-correction tasks.

In future work, we shall investigate why workers sometimes change their answers to emulate the self-correction reference answer. Clarifying the factors that lead workers to revise their answers will contribute to many applications. In addition, we will modify the experiment by changing settings such as the question difficulty and the number of times a question can be self-corrected. Another important future work is to find ways to measure the learning potential of workers, which can be an interesting human factor (Amer-Yahia and Roy 2016) in crowdsourcing.

Acknowledgments

We are grateful to Professor Hideaki Kuzuoka for helpful comments. This work was supported by JST CREST Grant Number JPMJCR16E3, Japan.

References

Abad, A.; Nabi, M.; and Moschitti, A. 2017. Autonomous crowdsourcing through human-machine collaborative learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, 873–876. New York, NY, USA: ACM.

Amer-Yahia, S., and Roy, S. B. 2016. Toward worker-centric crowdsourcing. *IEEE Data Eng. Bull.* 39(4):3–13.

Aroyo, L., and Welty, C. 2013. Measuring crowd truth for medical relation extraction. In *AAAI 2013 Fall Symposium on Semantics for Big Data*.

Ashikawa, M.; Kawamura, T.; and Ohsuga, A. 2015. Proposal of grade training method in private crowdsourcing system. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2017)*. ACM Association for Computing Machinery.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.* 51(1):7:1–7:40.

Das Sarma, A.; Parameswaran, A.; and Widom, J. 2016. Towards globally optimal crowdsourcing quality management: The uniform worker setting. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, 47–62. New York, NY, USA: ACM.

Doroudi, S.; Kamar, E.; Brunskill, E.; and Horvitz, E. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2623–2634. ACM.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1013–1022. ACM.

Drapeau, R.; Chilton, L. B.; Bragg, J.; and Weld, D. S. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640. ACM.

Gadiraju, U.; Fetahu, B.; Kawase, R.; Siehndel, P.; and Dietze, S. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24(4):30.

Gibson, J. J., and Gibson, E. J. 1955. Perceptual learning: Differentiation or enrichment? *Psychological review* 62(1):32.

Gibson, E. J. 1969. Principles of perceptual learning and development.

Haas, D.; Ansel, J.; Gu, L.; and Marcus, A. 2015. Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment* 8(12):1642–1653.

Hsieh, G., and Kocielnik, R. 2016. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 823–835. ACM.

Hung, N. Q. V.; Tam, N. T.; Lam, N. T.; and Aberer, K. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part II*, 1–15.

Hung, N. Q. V.; Thang, D. C.; Weidlich, M.; and Aberer, K. 2015. Minimizing efforts in validating crowd answers. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 999–1014. ACM.

Jagabathula, S.; Subramanian, L.; and Venkataraman, A. 2014. Reputation-based worker filtering in crowdsourcing. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2492–2500.

Joglekar, M.; Garcia-Molina, H.; and Parameswaran, A. 2013. Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, 686–694. New York, NY, USA: ACM.

Jung, H. J., and Lease, M. 2015. Modeling temporal crowd work quality with limited supervision. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Kinnaird, P.; Dabbish, L.; Kiesler, S.; and Faste, H. 2013. Co-worker transparency in a microtask marketplace. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1285–1290. ACM.

Law, E.; Yin, M.; Goh, J.; Chen, K.; Terry, M. A.; and Gajos, K. Z. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 4098–4110. New York, NY, USA: ACM.

Mettler, E., and Kellman, P. J. 2014. Adaptive response-time-based category sequencing in perceptual learning. *Vision research* 99:111–123.

Shah, N., and Zhou, D. 2016. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. In Balcan, M. F., and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1–10. New York, New York, USA: PMLR.

Suzuki, R.; Salehi, N.; Lam, M. S.; Marroquin, J. C.; and Bernstein, M. S. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 2645–2656. New York, NY, USA: ACM.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.