Human Computation (2021) 8:1:24

ISSN: 2330-8001, DOI: 10.15346/hc.v8i1.1

Empirical Study on Effects of Self-Correction in Crowdsourced Image Classification Tasks*

MASAKI KOBAYASHI, UNIVERSITY OF TSUKUBA

HIROMI MORITA, UNIVERSITY OF TSUKUBA

MASAKI MATSUBARA, UNIVERSITY OF TSUKUBA

NOBUYUKI SHIMIZU, YAHOO! JAPAN

ATSUYUKI MORISHIMA, UNIVERSITY OF TSUKUBA

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 their results according to the review. Self-correction was proposed as a complementary approach to statistical algorithms, in which workers independently perform the same task. It can provide higher-quality results with low additional costs. However, thus far, the effects have only been demonstrated in simulations, and empirical evaluations are required. In addition, as self-correction provides 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. We found that: (1) Self-correction is effective for making workers reconsider their judgments. (2) Self-correction is effective more if workers are shown the task results of higher-quality workers during the second stage. (3) A perceptual learning effect is observed in some cases. Self-correction can provide feedback that shows workers how to provide high-quality answers in future tasks. (4) A Perceptual learning effect is observed, particularly with workers who moderately change answers in the second stage. This suggests the possibility that we can estimate the learning potential of workers. These findings imply that requesters/crowdsourcing services can construct a positive loop for improved task results by the self-correction approach. However, (5) no long-term effects of the self-correction task are transferred to other similar tasks in two different settings.

*This paper is an extended version of a paper published at HCOMP 2018.

1. INTRODUCTION

Ensuring the quality of obtained data is a primary problem in crowdsourcing; numerous studies have attempted to improve 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 is to assign the same task to multiple workers and aggregate the results, which allows the final results to be computed through various aggregation methods (e.g., majority voting).

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 the task results of other workers to each worker after the results are submitted, allowing the worker to update their 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 the result-aggregation methods, in which multiple workers independently perform the same task, and the results are aggregated. Thus, self-correction provides better quality with little additional costs. However, the effectiveness of self-correction has only been shown in simulations, and real-world experiments have not been conducted. Therefore, its effectiveness in real-world settings is an interesting topic 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 the 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 the ability of workers to perform visual categorization tasks.

This paper reports the results of experiments designed to explore the effects of self-correction in a real crowdsourcing setting (Figure 1). In Experiment 1, we examine the short-term effect of the self-correction that improves the task result quality, and the long-term effect that improves the performance of workers by repeating self-correction tasks. In Experiment 2, we observe the short- and long-term effects with more difficult tasks than Experiment 1. In addition, we analyze the behavior of workers to distinguish workers who have the potential to improve their performance. In Experiment 3, we examine whether the long-term effects transfer to similar (but other) tasks or not. Our key findings are as follows:

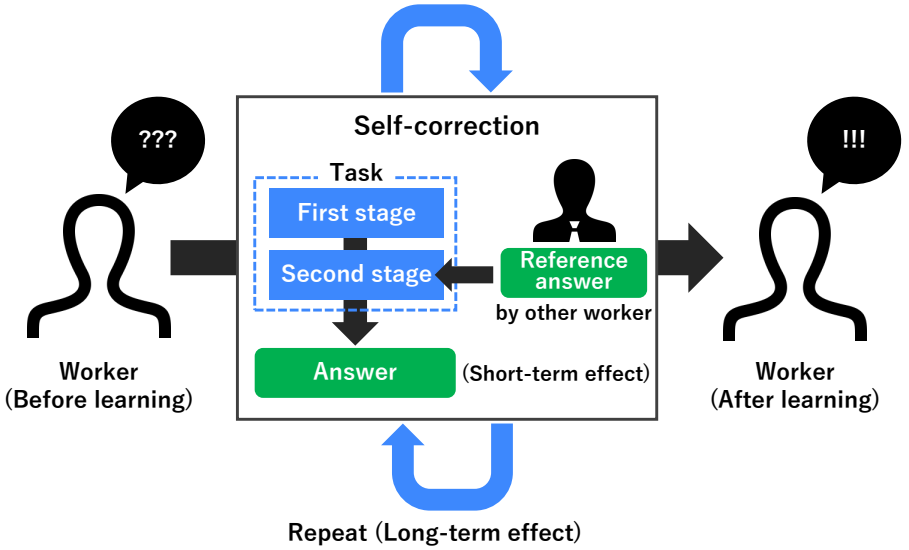

Figure 1. ***After workers answer the task question (first stage) with self-correction, they review the answers of other workers (second stage) and reconsider their answers in light of these reviews. In this study, we evaluate self-correction by workers, focusing on both short- and long-term effects.***

i. The short-term effects of self-correction on data quality were observed in a real-world setting. Here, the short-term effects were data quality improvements in the task. That is, workers notice the mistakes they made in the first stage. We found that the quality of data shown to workers in the second stage is important. Indeed, self-correction is more effective when workers were shown the task results answered by higher-quality workers in the second stage compared to when they simply reviewed their results.

ii. The long-term effects of self-correction on data quality were also observed. Here, the long-term effects involved perceptual learning by workers, namely quality improvements in a successive sequence of similar but different tasks. This result suggests that self-correction enables workers how to give high-quality answers during similar subsequent tasks.

iii. The long-term effects of self-correction were observed, particularly with the workers who moderately changed their answers in the second stage. To rephrase, these are the ones with moderate answer change rates (not too high or too low). This result suggests that we can estimate the learning potential of workers based on their behavior in self-correction tasks.

iv. Long-term effects of the self-correction tasks did not transfer to similar but other types of classification tasks. The perceptual learning performed here was only effective for the types of images classified during the learning phase.

All our findings were obtained using a simple monetary incentive. Each worker was paid according to the number of tasks performed. In (Shah and Zhou, 2016), the authors designed a monetary incentive that was proven to be theoretically optimal. Notably, our results showed that self-correction

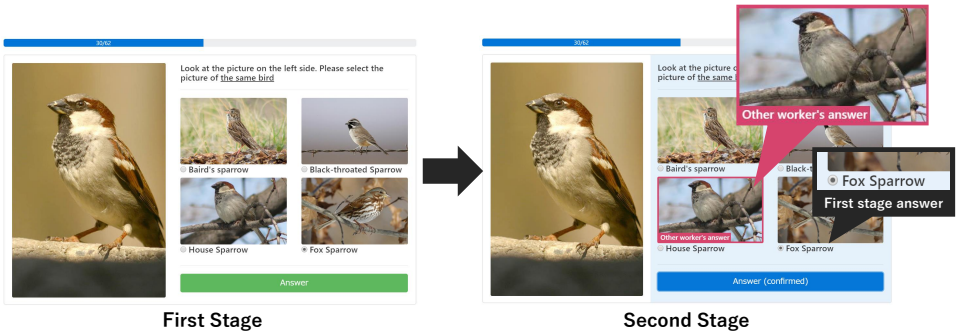

Figure 2. ***Self-correction tasks: In the first stage, a worker answers a question. In the second stage, the worker corrects their answer by reviewing other workers' answers.***

effectively improves the quality of both task results and workers, even with this simple monetary incentive. Introducing a more sophisticated incentive is one of the future work. However, such an incentive may limit the crowdsourcing services we can use with self-correction.

2. RELATED WORK

Improving the performance of workers, and thus the quality of data, is the area of focus in crowdsourcing study, and numerous studies have addressed these issues (Daniel et al., 2018). New advances within this area will accelerate human and machine collaboration, like data-driven machine learning techniques (Vaughan, 2018)(Zhang et al., 2016)(Yan et al., 2011).

One dominant 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 various factors, such as the quality of the workers, agreement ratios, and clustering results (Hung et al., 2013) (Jagabathula et al., 2014) (Aroyo and Welty, 2013). (Quoc Viet Hung et al., 2013) proposed a benchmarking framework to evaluate aggregation methods in crowdsourcing. Other approaches include those addressing better task design (Doroudi et al., 2016) and incentive structures (Kinnaird et al., 2013) (Hsieh and Kocielnik, 2016) (Matsubara and Wang, 2014). 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 et al., 2017) and Microtalk (Drapeau et al., 2016) give workers opportunities to change their answers after seeing the justifications for answers of other workers. (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, how this feedback works from a scientific perspective has not been studied thus far.

It is known that assessments by workers are biased. (Gadiraju et al., 2017) showed that crowd workers often lack awareness about their actual level of competence. Incorporating such bias into a self-correction framework would be a 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 regular tasks (Ashikawa et al., 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., 2016) proposed the micro-internships framework that connects interns to experts through crowdsourcing for helping them in obtaining the required skills. Crowd Coach is a skills development system that conducts peer coaching while performing tasks (Chiang et al., 2018).

Estimating the performance from the behavior/output of workers is useful to ensure the quality of crowdsourcing (Rzeszotarski and Kittur, 2011) (Rzeszotarski and Kittur, 2012). (Oyama et al., 2013) proposed a method to aggregate task results based on self-reported confidence scores from the workers. (Lebreton et al., 2015) compared self-reported metrics and actual worker behavior using eye-tracking techniques with local webcams of workers. Some research show techniques for measuring worker quality without the use of so-called gold standard data (Joglekar et al., 2013) (Das Sarma et al., 2016). In this paper, we also analyzed the behavior of workers on self-correction tasks, which we refer to as answer change patterns. Sensitive information, such as the personal behavior information of workers, must be strictly protected (Xia et al., 2017).

It is expected that perceptual learning occurs in microtasks with feedbacks. (Abad et al., 2017) showed that rule-based feedback given to workers who provided incorrect answers is useful in training. 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 essential key to perceptual learning is how many times each worker performs the same tasks. (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 in our framework in the future.

In this paper, we show that, with self-correction, more evident improvements are shown in the quality of task results (i.e., short-term effects) by high-quality workers in the second assignment. However, the self-correction does not mention how to find high-quality workers 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 most straightforward approach is to insert particular tasks into the sequence of regular tasks for measuring worker quality. (Jung and Lease, 2015) showed an approach that measuring workers by limited gold standard questions. (Hata et al., 2017) showed that the ability of workers could be estimated by the first five tasks for long-term the quality of workers. However, to cause long-term effects and observe behaviors during tasks, workers must work on more number of tasks. There are risks that workers leave from tasks in the middle with no rewards (Han et al., 2019).

3. SELF-CORRECTION IN MICROTASKS

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

Table 1. ***Experiment settings***

	Experiment 1	Experiment 2	Experiment 3
Group Conditions	Trusted vs. Self	Correct vs. Random	Correct vs. No self-correction
Task Difficulty	Easy	Difficult	Easy & Difficult
Dataset in Learning and Testing	Same	Same	Different
Learning Phase	Twice, 28 tasks	Twice, 52 tasks + 2 gold standard questions	Once, 48 + 4 gold standard questions
Image Dataset	Caltech-UCSD Birds 200	Abstract paintings from wikiart.org	Caltech-UCSD Birds 200
Filter	Under 25% in mid or post phase	Gold standard questions	Gold standard questions

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 the first answer. Then, during the second stage, the worker can revise this answer after reviewing answers of other workers.

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. We did not adopt the reward system for our experiments because our experiments focus on the task design (two-stage setting with feedback) for self-correction.

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

4. EXPERIMENT 1: REFERENCE SOURCE COMPARISON (TRUSTED VS SELF)

4.1. Purpose of the experiment

We conducted Experiment 1 to investigate the following 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

Table 2. Procedure of Experiment 1

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

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.

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 approximately $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 answers of other workers as a reference for the trusted group or without the reference for the self group.

4.3. Procedure

Workers were asked to perform three phases of test tasks and two phases of self-correction tasks, listed 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 clarified the involuntary learning effects of self-correction tasks, i.e. long-term effects. 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, i.e. short-term effect.

4.4. Tasks

Each Yahoo! Crowdsourcing task consisted of a number of classification tasks, such as shown in Table 1. The classification tasks were generated by using Crowd4U [2].

4.4.1. Test phase and the first stage of the self-correction phase

In the test phase and the 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

[1] https://crowdsourcing.yahoo.co.jp

[2] https://crowd4u.org

Table 3. ***Pre-test phase accuracy and overall working time (sec) in Experiment 1***

			Pre-test Accuracy			Overall Working Time		
Condition	Filter	N	Median	Mean	Std	Median	Mean	Std
self	None	98	0.833	0.826	0.147	487.07	538.19	193.67
	Under 25%	84	0.833	0.83	0.136	517.44	553.74	202.05
trusted	None	98	0.833	0.816	0.131	526.01	549.60	175.88
	Under 25%	86	0.833	0.824	0.134	530.47	563.17	179.90

particular image. We displayed an image of a bird to be classified 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 sparrow family, and classified into four genus, namely baird's sparrow, black-throated sparrow, house sparrow, and fox sparrow. Throughout 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.

4.4.2. Second stage of self-correction phase

In the second stage of the self-correction phase, we presented the same image and photos as the first stage with the first stage choice marked. Workers could change their choice. We highlighted answers of other workers as reference answers for workers in trusted group (Right side in Figure 2).

4.5. Group conditions

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 with accuracy rate below 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 (Table 3).

4.7. Results

Table 3 shows the accuracy in the pre-test phase and the overall working time. The workers spent about nine minutes on average; in addition, the hourly rate was approximately 6.6 USD.

4.7.1. 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 ($F(1, 168) = 40.36$, $p < .001$), and from the presence of reference answer ($F(1, 168) = 10.45$, $p < .001$) and their interaction ($F(1, 168) = 48.29$, $p < .001$).

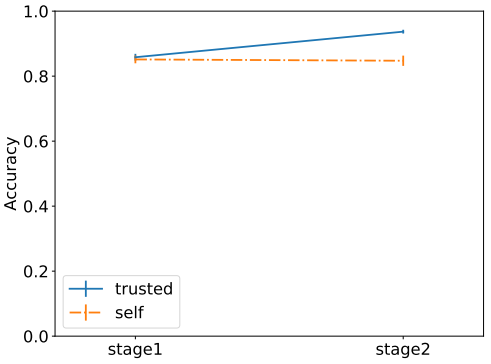

Figure 3. *[Experiment 1] Accuracy rate for the first and second stages of each learning phase.*

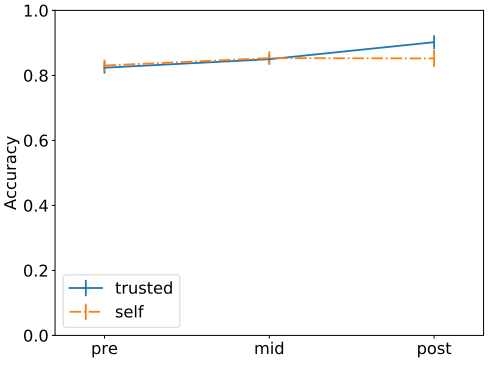

Figure 4. *[Experiment 1] Accuracy rate for each test phase.*

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" ($F(1, 85) = 54.84$, $p < .001$), but no main effect from the stage for workers in the "self group" ($F(1, 83) = 0.61$, n.s.). There was no simple main effect from the presence of reference answers for the first stage ($F(1, 168) = 0.175$, n.s.), but there was a simple main effect from the presence of reference answers for the second stage ($F(1, 168) = 31.82$, $p < .001$).

There was no difference in the accuracy rate of the first stage answers between worker groups and 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 answers other workers is an important factor in increasing the accuracy rate of the second stage answers and the self-correction strategy can improve the quality of the output of the workers if given an appropriate reference answer.

4.7.2. 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 presence of reference answer as a between-worker factor. The ANOVA revealed a significant effect from the test phase ($F(2, 336) = 8.73$, $p < .001$), but no effect from the type of reference answer ($F(1, 168) = 0.64$, n.s.).

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. They showed a simple main effect from the test phase for workers in the "trusted group" ($F(2, 170) = 11.82$, $p < .001$), but no simple main effect from the test phase for workers in the "self group" ($F(2, 166) = 1.08$, n.s.). A multiple-comparison using a Bonferroni correction with the correct answer rates for the "trusted group" revealed no difference between the pre-test and mid-test rates, 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 ($F(1, 168) = 0.105$, n.s.; $F(1, 168) = 0.027$, n.s.) ; however, there was a simple main effect for the post-test ($F(1, 168) = 4.48$, $p < .05$).

Workers in the trusted group increased the accuracy of answers from pre-test to post-test, while workers in the self group did not. As a result, a difference in the accuracy between the two groups at the post-test was evident, although workers of two groups did not differ in ability at first. Thus, it was proven that workers developed an ability to give high-quality answers by repeating self-correction tasks with appropriate reference answers.

4.7.3. Reaction time and accuracy

We evaluated the correlation between the accuracy of stages and reaction time. The analysis shows that no correlation was observed in stage 1 and stage 2 in the trusted condition (stage 1: r=-0.067, stage 2: r=-0.047). In addition, correlation was not found in self condition (stage 1: r=0.085, stage 2: r=-0.050). Further, we evaluated the correlation between the test accuracy and reaction time. The analysis shows that correlation was not observed in pre- and post-test in trusted condition (pre: r=0.151, post: r=-0.047). In addition, no correlation was observed in pre- and post-test even in self condition (pre: r=-0.189, post: r=-0.016). Overall, we did not observe correlations between accuracies and reaction time. In this experiment, the average accuracy was high, and the data that could be judged instantly were used. It seems that the influence of the experimental conditions on the reaction time was minimal.

5. EXPERIMENT 2: REFERENCE RELIABILITY COMPARISON (CORRECT VS RANDOM)

5.1. Purpose of the experiment

We conducted Experiment 2 to investigate the following questions: (1) Does the self-correction strategy work well with more difficult tasks? (Short-term effect) (2) Are involuntary learning effects observed with more difficult tasks? (Long-term effect) (3) Is it possible to distinguish, without gold standard questions, workers who will improve their performance from those who will not improve?

5.2. Participants

One-hundred ninety-one workers participated in the 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

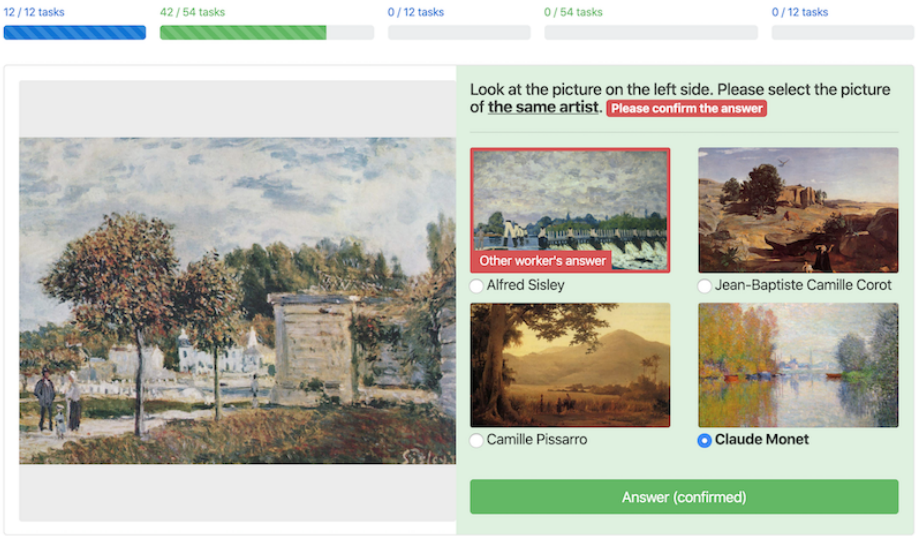

Figure 5. ***Second stage of the self-correction task for Experiment 2. Workers were asked to classify paintings by impressionist artist.***

self-correction with a random answer (hereafter, "random"). The workers were to receive a reward of approximately $1 when they completed all set of tasks.

5.3. Procedure

This experiment followed the same procedure as that of Experiment 1, however here, each learning phase contained 52 self-correction tasks. We provided workers with more learning opportunities due to the increased difficulty of this task (Table 2).

5.4. Tasks

The same tasks as used in Experiment 1 were used, with the difference that workers classified paintings instead of birds (Figure 5). Pieces of art for the tasks were selected from the work of four well-known impressionist painters: (1) Alfred Sisley, (2) Jean-Baptiste Camille Corot, (3) Camille Pissarro, and (4) Claude Monet. We collected painting images from wikiart.org[3].

5.5. Group conditions

The correct answers were used as appropriate reference answers in the second stage. We did not use trusted worker's answers as the reference answer due to the task being so difficult that very few workers were expected to be able to give fully trusted answers. Answers randomly selected from four choices were used as inappropriate reference answers.

5.6. Answer change rate

The part of the analysis of this experiment focused on how workers changed their answers during a self-correction task. Here, T is the set of self-correction task results of a worker. Every $t \in T$ has a

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

Table 4. ***Pre-test phase accuracy and overall working time (sec) in Experiment 2***

			Pre-test Accuracy			Overall Working Time		
Condition	Filter	N	Median	Mean	Std	Median	Mean	Std
random	None	105	0.333	0.354	0.152	823.79	902.76	482.83
	Gold	99	0.333	0.357	0.153	832.65	923.78	473.50
correct	None	86	0.333	0.356	0.15	876.06	898.56	379.89
	Gold	82	0.333	0.363	0.153	885.90	914.27	373.76

(stage1, stage2) answer pair. When the answer at stage2 is different from that at stage1, the worker changed the answer at stage2. We calculate answer change rate of the worker by equation (1).

$$answer\_change\_rate = \frac{|\{t|t_{stage1} \neq t_{stage2}, t \in T\}|}{|T|} \tag{1}$$

5.7. Filter

We added a gold-standard question to the task list to identify and omit workers who might have performed tasks inadequately due to fatigue or satisfice. In the gold-standard question, one of the four exemplar paintings was presented to be classified. We analyzed only those workers who correctly answered at least two of four gold-standard questions.

5.8. Results

We analyzed task results from 181 workers who passed tasks for filtering. Table 4 shows the accuracy in the pre-test phase and the overall working time. The workers spent approximately fifteen minutes on average, and the hourly rate was approximately 4.0 USD. The average accuracy was lower than Experiment 1 because relatively difficult tasks were prepared.

5.8.1. 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 stages as a within-worker factor, and the quality of reference answers (correct vs random) as a between-worker factor. There were main effects of the stages ($F(1, 179) = 83.97$, $p < .001$) and the quality of reference answers ($F(1, 179) = 52.34$, $p < .001$). There was a significant interaction between the stages and the quality of reference answers ($F(1, 179) = 105.27$, $p < .001$). 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 ($F(1, 81) = 91.05$, $p < .001$); however, there was no simple main effect from the stage for workers in the random group ($F(1, 98) = 0.15$, n.s.). There was no simple main effect from the quality of reference answers for the first stage ($F(1, 179) = 1.79$, n.s.) ; however, there was a simple main effect from the quality of reference answers for the second stage ($F(1, 179) = 91.32$, $p < .001$).

The results supported the claim that has been shown in Experiment 1, because presenting the correct answers in the second stage did improve the results, but presenting random answers did not. However, note that some workers might have always adopted the reference answer in the second

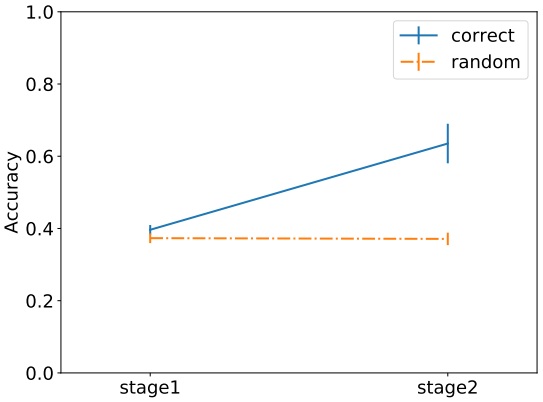

Figure 6. ***[Experiment 2] Accuracy rate of the first and second stages of the learning phases.***

stage and obtained good results in the correct-reference condition. They will affect the growth rate, which will be described later.

5.8.2. Long-term effect

Figure 7 shows the accuracy rate of the test phases of each group.

We conducted a two-way ANOVA with the accuracy rate as a dependent variable, the test phases as a within-worker factor, and the quality of reference answers (correct vs random) as a between-worker factor. There was a main effect of the test phases ($F(2, 358) = 3.89$, $p < .05$) but no main effect of the quality of reference answers ($F(1, 179) = 2.016$, n.s.).There was no significant interaction between the test phases and the quality of reference answers ($F(2, 358) = 1.61$, 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 the accuracy rates for the correct-reference group revealed a significant difference between the pre-test and post-test ($p < .05$), no significant differences between the pre-test and mid-test and between the mid-test and post-test. A multiple-comparison with accuracy rates for the random-reference group showed no significant difference between the pre-test and mid-test, the pre-test and post-test, and the mid-test and post-test.

The accuracy rate of the correct-reference group slightly improved between the pre-test and post-test. In contrast, the accuracy rate of the random-reference group did not differ between pre-test and post-test. The pattern of the results overall supports that the long-term effect of involuntary learning can be seen even in difficult tasks. The effect was slight but significant.

5.8.3. Answer change rate and tests

The long-term learning effect was small. However, it is assumed that there would be a difference in the amount of the learning effect between workers who were motivated for the tasks and those who were not, and it is expected that workers who were motivated would have learned more than those who were not.

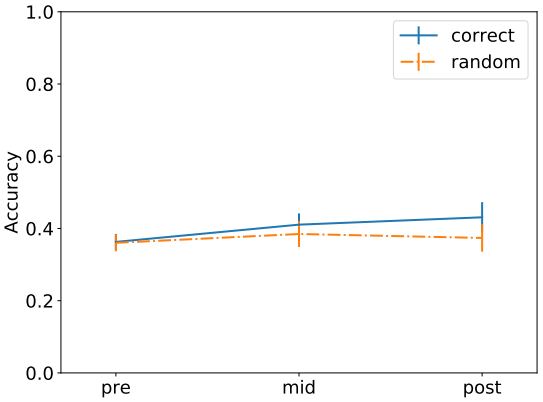

Figure 7. *[Experiment 2] Accuracy rate and test phases for each group condition.*

Table 5. *Answer change rate groups for Experiment 2*

	condition	
answer change rate	correct	random
$0.0 \sim 0.2$	41	65
$0.2 \sim 0.4$	21	14
$0.4 \sim 0.6$	11	9
$0.6 \sim 0.8$	9	9
$0.8 \sim 1.0$	0	2

We hypothesized that worker motivation appears in the rate with which they changed their first stage answer in the second stage. Motivated workers will consider carefully the reference answer and take advantage of the opportunity to change their answers. On the other hand, workers who ignore the reference answer and stick to their first stage answer or workers who always follow the reference answer are considered to be less motivated. Thus we compared the learning effect between the group of different answer change rate. Table 5 shows the distribution of groups of the answer change rate for each condition. We focused on 4 groups, excluding 0.8-1.0, for statistical analysis because that group contains few workers.

Figure 8 shows the accuracy of a test phase in each group of answer change rate.

For workers in the correct group, we conducted a two-way ANOVA with the accuracy rate of tests as a dependent variable, the test phase as a within-worker factor and the group of answer change rate as a between-worker factor. There was a no main effect of the test phase ($F(2, 156) = 2.07$, n.s.) but there was a main effect from the answer-change-rate group ($F(3, 78) = 5.88$, $p < .005$) and interaction between the test phase and group ($F(6, 156) = 2.26$, $p < .05$). The post-hoc analysis revealed a simple main effect from test phase only for the group 2 (answer change rate of 0.2 - 0.4, $p < .005$) and multiple comparison yielded a significant difference between pre-test and post-test

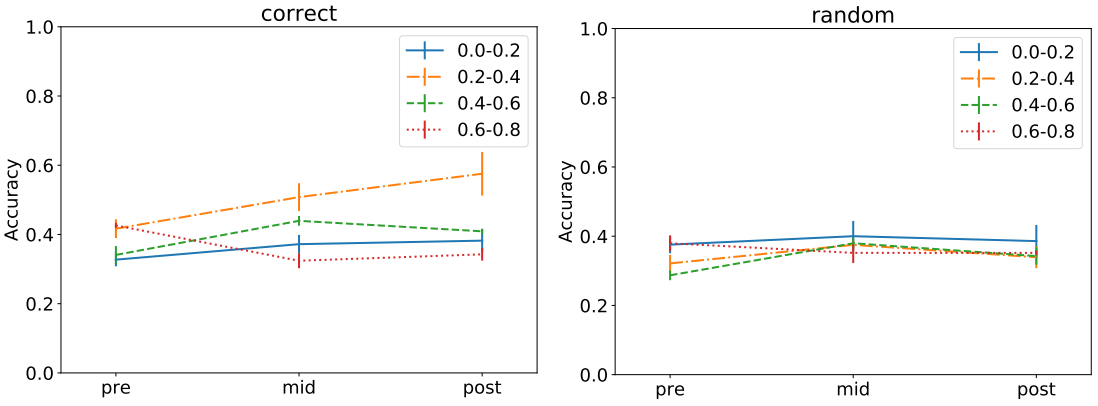

Figure 8. ***Answer change rate groups and test phases.***

($p < .01$). The post-hoc analysis also revealed a simple main effect from answer-change-rate-group for mid- and post-test ($F(3, 78) = 4.00$, $p < .05$; $F(3, 78) = 5.70$, $p < .005$) but not for pre-test. Multiple comparison yielded significant differences between pairs of group1 (answer change rate of 0 - 0.2) and group2 ($p < .05$), and group2 and group4 ($p < .05$) for mid-test. For post-test, multiple comparison yielded significant differences between group1 and group2 ($p < .05$) and group2 and group4 ($p < .05$).

For workers in the random 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 answer-change-rate group as a between-worker factor. There was no main effect from test phase or group($F(2, 186) = 0.74$, n.s.; $F(3, 93) = 0.69$, n.s.). Interaction of them was not significant ($F(6, 186) = 0.29$, n.s.).

The improvement in the accuracy rate of the worker's test task (i.e., the long-term effect) was only seen in the correct condition group with an answer change rate of 0.2 to 0.4. Their accuracy rate increased from 0.42 to 0.58. This result suggests that the self-correction task does not improve the performance of all workers, but is particularly effective for a specific range of workers.

It was suggested that by looking at the answer change rate of workers who worked on appropriate self-correction tasks, workers with the performance expected to improve in the future can be identified. It is essential to provide appropriate motivation and incentives so that workers can continue to work on self-correction tasks.

The answer change rate group to which the excellent workers belong is expected to vary depending on the nature and difficulty of the task the worker is working on.

5.8.4. Reaction time and accuracy

We evaluated the correlation between the accuracy of stages and reaction time. In the correct condition, a weak and moderate correlation was observed in stage 1 and stage 2, respectively (stage 1: r=0.331, stage 2: r=0.419). In the random condition, a moderate correlation was observed in stage 1 (r=0.412); however, no correlation was observed in stage 2 (r=0.145). In addition, we evaluated

the correlation between the test accuracy and reaction time. In the trusted condition, no correlation was observed in the pre-test (r=-0.052); however, a weak correlation was observed in the post-test (r=0.383). In the random condition, no correlation was observed in the pre-test (r=0.061); however, a weak correlation was observed in the post-test (r=0.321).

Experiment 2 used a more difficult and abstract task than Experiment 1; therefore, weak relationships were observed between the reaction time and accuracy. In the correct condition of experiment 2, compared with random, answers of workers and the reference answer are often different because the tasks are difficult. Thus, they seem to compare them carefully.

6. EXPERIMENT 3: DATASET COMPARISON (SAME VS DIFFERENT)

6.1. Purpose of the experiment

The next question is whether there are transfer effects in the self-correction tasks. In Experiment 1 and 2, we used the same classes of images in both of the learning and test phases. In contrast, this experiment uses a different set of images in the test phases from that in the learning phase and checks whether there is a long-term effect, i.e., whether the learning phase with a set of images improves the result with a different set of images in the test phase.

6.2. Participants

One-hundred ninety-one workers participated in the 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, "SC"). The other half was assigned to a group engaging in one step tasks like test phase (hereafter, "NSC"). The workers were to receive a reward of about $1 when they completed all the tasks.

6.3. Procedure

Table 6 shows the procedure of the experiment. This procedure is slightly different from those for Experiments 1 and 2 for the following reasons. Firstly, since we emphasize whether transfer effects can be observed, we do not focus on how they happen. Therefore, this experiment has only one learning phase. Secondly, we expected that the learning effect would be smaller than those found in Experiments 1 and 2. Therefore, we doubled the number of tasks to 24 in the test phases to catch the effects.

6.4. Tasks

The same task as used in Experiment 1 was used here (Figure 9), with the difference that workers were asked to classify different classes of birds for the learning phase and the test phase.

Obtaining Two Set of Tasks. For the experiment, we used two different sets of tasks to check whether there were any transfer effects. It was necessary to balance the difficulty of the set of tasks for learning and that for the tests.

For this purpose, we asked workers in a different platform (Amazon's Mechanical Turk) to perform the classification tasks with several data sets. Then we selected a pair of data sets (Dataset 1 and Dataset 2) for which we observed the same average accuracy of 50% and another pair of data sets (Dataset 3 and Dataset 4) for which we observed the same average accuracy of 90% in performing the tasks. We assume that the data sets in each paper would result in the same accuracy with workers in Yahoo! Crowdsourcing.

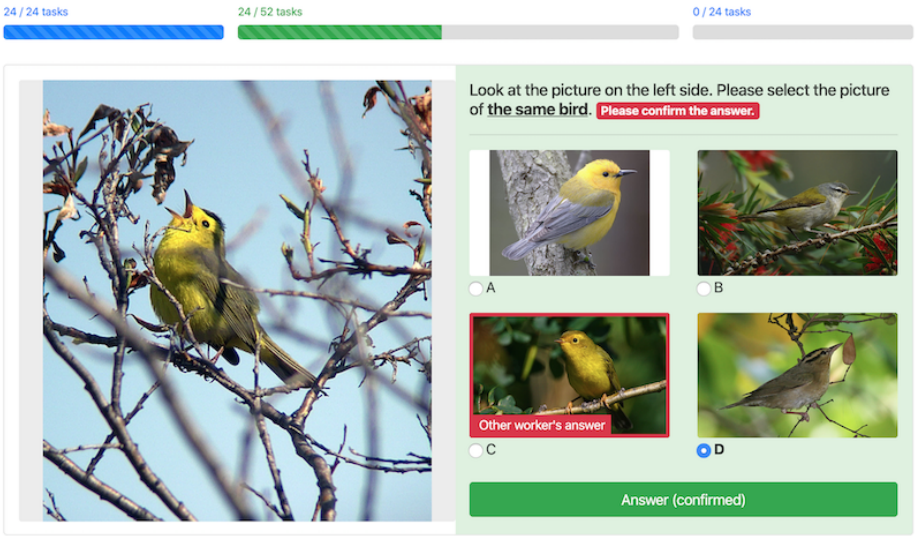

Figure 9. ***The second stage of self-correction task for Experiment 3.***

Table 6. ***Procedure of Experiment 3***

	Phase	Task type	Number of tasks
1	Pre-test	Test	24
2	Learning 1	Self-correction	48 + 4 gold tasks
3	Post-test	Test	24

6.5. Group Conditions

We compared a group (named 'SC') of workers who performed Self-Correction tasks with correct answers and another group (named 'NSC') of workers who performed tasks that are not Self-correction, in the accuracy of task results.

6.6. Filter

We applied the same filtering method to the data as the one used in Experiment 2. We analyzed only those workers who correctly answered at least two of four gold-standard questions.

6.7. Results

Table 7 shows the accuracy in the pre-test phase and the overall working time. The workers spent approximately eleven minutes on average, and the hourly rate was approximately 5.6 USD. Notably, sixty-six of the workers in Experiment 1 also participated in Experiment 3. Experiment 3 was conducted 18 months later after Experiment 1. We assume the experience in experiment 1 did not affect the results of Experiment 3.

Figure 10 shows the results of each data set and conditions. The x-axis shows test phases, and y-axis shows the accuracy.

Table 7. ***Pre-test phase accuracy and overall working time (sec) in Experiment 3***

				Pre-test Accuracy			Overall Working Time		
Dataset Pair	Condition	Filter	N	Median	Mean	Std	Median	Mean	Std
1, 2	NSC	None	43	0.375	0.381	0.128	693.33	733.11	364.89
(learning, test)		Gold	40	0.375	0.372	0.124	690.11	715.81	339.92
	SC	None	46	0.333	0.348	0.14	665.79	733.05	370.57
		Gold	42	0.417	0.354	0.145	674.32	751.56	358.67
2, 1	NSC	None	59	0.375	0.386	0.129	579.96	653.04	315.65
		Gold	56	0.375	0.395	0.125	588.99	658.82	304.44
	SC	None	44	0.417	0.447	0.106	710.77	796.44	339.48
		Gold	41	0.417	0.454	0.105	743.12	811.40	346.99
3, 4	NSC	None	49	0.75	0.736	0.152	445.94	482.46	173.15
		Gold	48	0.75	0.744	0.144	456.19	487.30	171.60
	SC	None	49	0.792	0.758	0.166	501.64	547.43	169.19
		Gold	48	0.792	0.767	0.153	514.98	552.27	167.51
4, 3	NSC	None	51	0.917	0.902	0.106	440.74	502.14	217.93
		Gold	51	0.917	0.902	0.106	440.74	502.14	217.93
	SC	None	51	0.917	0.895	0.091	581.45	658.53	288.87
		Gold	51	0.917	0.895	0.091	581.45	658.53	288.87

We conducted two-way ANOVA for each dataset with the accuracy rate as a dependent variable, the test phase as a within-worker factor, and the condition of task type (correct vs. non-SC) as a between-worker factor. As a result, in dataset2, there was a significant effect from the test phase ($F(1, 91) = 4.838$, $p < .05$), but there were no significant effect from the task type ($F(1, 91) = 2.571$, n.s.) and the interaction ($F(1, 91) = .019$, n.s). In dataset0, dataset1, and dataset3, there were no significant effects from the test phase ($F(1, 63) = 0.214$, n.s.; $F(1, 68) = 0.947$, n.s.; $F(1, 77) = .09$, n.s.), the task type ($F(1, 63) = .162$, n.s.; $F(1, 68) = 1.082$, n.s.; $F(1, 77) = 3.327$, n.s.), and the interaction ($F(1, 63) = .049$, n.s.; $F(1, 68) = 2.630$, n.s.; $F(1, 77) = .165$, n.s.).

The accuracy rate of the workers did not increase in the post-test with data0, data1, and data3. It increased with data2. However, since the accuracy rate increased for both of correct group and non-SC group, it is suggested that the improvement in accuracy is likely due to workers growing familiarization with the tasks. Therefore, no evidence was obtained which proved that long-term effects transfer to the classification tasks with different types of data sets.

7. DISCUSSION

7.1. 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 misjudgment commonly observed in crowdsourcing, if an elaborate incentive mechanism is used. In our experiment, real crowd workers performed tasks with self-correction. The

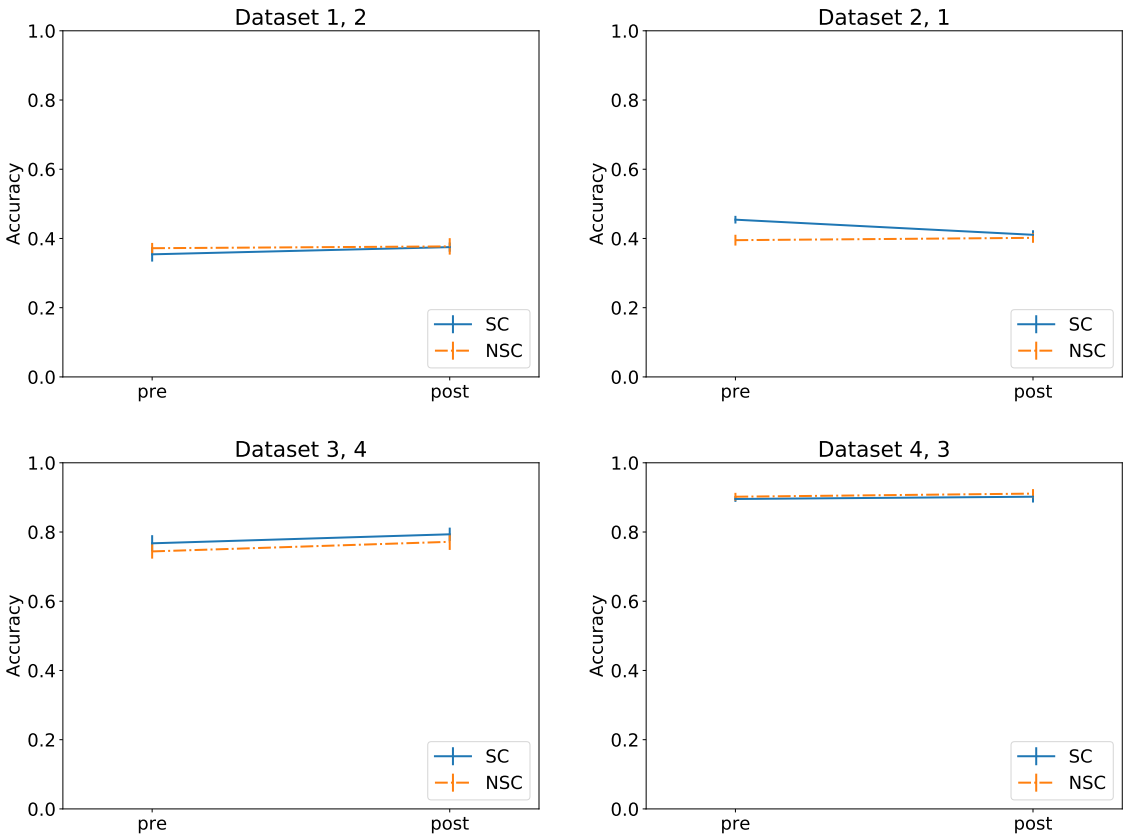

Figure 10. ***Accuracy rate with each test phase for each dataset.***

experimental results revealed that the task results obtained with self-correction are more accurate than without. Thus, for answers that are more accurate, it is important to provide answers of other workers 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 the 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% of 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.

Our main contribution is to show the short-term effect of self-correction tasks in the real-world task setting. The experiments were conducted with two datasets with very different characteristics (such as the level of abstraction, the target, the difficulty level); thus, we found that the short-term effects can be observed in both datasets. A more detailed exploration into its factors remains an important future work. However, reporting the results of these two datasets with classification tasks would be

a valuable first step in the investigation of the effectiveness of self-correction.

7.2. Long-term effect of self-correction

By comparing worker performance in the pre-test and post-test phases, we found a 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 the task result is one of the most important issues in crowdsourcing. Assigning additional tasks to them for training and then retaining learned workers is one way to support the issue. However, the training requires workers to attend training sessions 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 become experts while completing tasks and receiving rewards. This development will require a certain amount of time or 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 the accuracy rate for the workers in the "correct" group between the pre-test and post-test phases. From our experimental results, the relationship between the number of tasks required for learning and task difficulty is not clear. Although there was no difference in performance between the two groups ("correct" and "random') in the post-test phase, we expect the apparent difference to appear after performing more self-correction tasks.

In Experiment 1, we observed the long-term effects of repeating self-correction, but it was not clear whether all workers were receiving long-term effects or only some workers. In Experiment 2, we analyzed the behavior of workers in self-correction tasks as an answer change rate and analyzed what kind of workers have obtained a long-term effect through tasks.

In Experiment 3, we tested the hypothesis that the long-term effects are transferred to related other tasks, and the results showed the long-term effects were not transferred. We believe this result is scientifically valuable to show the limitation of self-correction tasks in real-world settings. However, an in-depth investigation of the conditions in which the learning effect is transferred remains future work. We assume these conditions contain (1) the number of assigned self-correction tasks, and (2) difficulty and relevance of datasets.

To find a good worker, it is important to find first 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 phase) by noting the answer patterns.

7.3. Relation between the answer change rate and the long-term effect

Is there a relationship between how seriously a worker considered the reference answer and the quality of that worker? To find out, we categorized the workers depending on their frequency of changing the second-stage answer from the first. As a result, workers presented with correct references who changed less than 20% or more than 40% hardly improved their accuracy from pre-test to post-test. In contrast, workers who changed their answer at the rate of 20% – 40% improved their quality of answers from the average accuracy rate of 0.42 to 0.6. No evident relation was found between the answer change rate and the improvement for the random condition.

To summarize, comparing to workers who hardly changed their first stage answer as well as those who frequently changed, those who had a modest rate of change eventually made relatively high-

quality output. Accordingly, there exists a possibility that we can distinguish workers who have a learning potential based on the rate of changed answers in the self-correction. To our knowledge, no previous study of self-correction relates the behavior of workers, such as change rate, to their quality. We expect that future studies can lead to post-hoc estimation of the motivation of workers or the ability to answer their tasks based on their behavior in self-correction.

7.4. Reaction time and accuracy

Weak correlations were observed between time spent and performance, especially in Experiment 2. We assume that the difficulty of tasks and the quality of the reference answer may explain this difference. In the correct condition of experiment 2, as compared with the random condition, answers of workers and the reference answer were often different because the tasks were difficult. Thus, they seem to compare them carefully. However, we would like to leave the detailed analysis of this issue for future work.

7.5. Incentives for workers

A reward algorithm for self-correction was proposed for workers who behaved seriously. 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 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.

7.6. Choosing workers for reference answers

In Experiment 1, we chose workers by the accuracy rate to provide reference answers in self-correction tasks. We used correct labels from an image dataset in Experiment 2. 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 tasks with known correct labels such as gold standard questions.

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 abilities of workers. Identifying the best or better combinations for finding better workers is an interesting research topic. Although addressing this issue is beyond the scope of this paper, we expect that the answer will depend on the nature of the tasks.

7.7. Deployment

There is more than one method to deploy self-correction for a set T of tasks to be submitted to typical commercial crowdsourcing platforms.

A simple framework can be configured as follows: First, ask the workers to perform a small set of test tasks (with gold-standard data) to measure their quality. Subsequently, select the top $X\%$

workers based on the quality (denoted by E). Then, assign a subset of T (referred to as a batch) to the E without the second stage to obtain answers to be used in the second stage for other workers (are not in E). Ask other workers to perform the same batch by self-correction tasks with Es labels. Finally, evaluate answers of the final $n\%$ 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.

This kind of framework would create a positive loop for accurate crowdsourcing results while providing skill development experiences for workers. We can consider combining dynamic reward algorithms, techniques to identify good workers, and other methods if the crowdsourcing platform allows it.

8. 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, in particular with the workers who moderately change the answers in the second stage. Self-correction can give feedback that shows workers how to provide high-quality answers in future tasks. This suggests the possibility that we can estimate the learning potential of workers.
- However, no long-term effects of the self-correction task were found to be transferred to other similar tasks.

The findings imply that requesters/crowdsourcing services can construct a positive feedback loop to improve the quality of workers effectively, given a homogeneous set of microtasks. Thus, we suggest a couple of interesting topics for future research: (1) Observe stronger learning effects in the long-term by working with a self-correction that requires breaks or intervals. (2) Clarify the important factors for establishing both worker learning and transfer to other tasks to leverage long-term effects for future tasks. (3) Encourage workers to make better decisions rather than simply rejecting workers based on their answer change rate.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR16E3, Japan.

9. 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)*. ACM, New York, NY, USA, 873–876. DOI:http://dx.doi.org/10.1145/3077136.3080666

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). (May 2017). https://www.microsoft.com/en-us/research/publication/revolt-collaborative-crowdsourcing-labeling-machine-learning-datasets/

Chiang, C, Kasunic, A, and Savage, S. (2018). Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *CoRR* abs/1811.05364 (2018). http://arxiv.org/abs/1811.05364

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, Article 7 (Jan. 2018), 40 pages. DOI: http://dx.doi.org/10.1145/3148148

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)*. ACM, New York, NY, USA, 47–62. DOI:http://dx.doi.org/10.1145/2882903.2882953

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*. ACM, 2623–2634.

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*. ACM, 1013–1022.

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, 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 (2017), 30.

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*. ACM, 1631–1640.

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

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

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

Han, L, Roitero, K, Gadiraju, U, Sarasua, C, Checco, A, Maddalena, E, and Demartini, G. (2019). All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 321–329.

Hata, K, Krishna, R, Fei-Fei, L, and Bernstein, M. S. (2017). A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 889–901. DOI:http://dx.doi.org/10.1145/2998181.2998248

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*. ACM, 823–835.

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. http://dx.doi.org/10.1007/978-3-642-41154-0_1

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*. ACM, 999–1014.

Jagabathula, S, Subramanian, L, and Venkataraman, A. (2014). Reputation-based Worker Filtering in Crowdsourcing. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2492–2500. http://papers.nips.cc/paper/5393-reputation-based-worker-filtering-in-crowdsourcing.pdf

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)*. ACM, New York, NY, USA, 686–694. DOI:http://dx.doi.org/10.1145/2487575.2487595

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*. ACM, 1285–1290.

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)*. ACM, New York, NY, USA, 4098–4110. http://doi.acm.org/10.1145/2858036.2858144

Lebreton, P, Hupont, I, Mäki, T, Skodras, E, and Hirth, M. (2015). Eye tracker in the wild: studying the delta between what is said and measured in a crowdsourcing experiment. In *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*. 3–8.

Matsubara, S and Wang, M. (2014). Preventing Participation of Insincere Workers in Crowdsourcing by Using Pay-for-Performance Payments. *IEICE Transactions on Information and Systems* E97.D, 9 (2014), 2415–2422. DOI:http://dx.doi.org/10.1587/transinf.

2013EDP7441

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

Oyama, S, Baba, Y, Sakurai, Y, and Kashima, H. (2013). Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *IJCAI 2013 - Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. 2554–2560.

Quoc Viet Hung, N, Tam, N. T, Tran, L. N, and Aberer, K. (2013). An Evaluation of Aggregation Techniques in Crowdsourcing. In *Web Information Systems Engineering – WISE 2013*, Xuemin Lin, Yannis Manolopoulos, Divesh Srivastava, and Guangyan Huang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

Rzeszotarski, J and Kittur, A. (2012). CrowdScape: Interactively Visualizing User Behavior and Output. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 55–62. DOI:http://dx.doi.org/10.1145/2380116.2380125

Rzeszotarski, J. M and Kittur, A. (2011). Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 13–22. DOI:http://dx.doi.org/10.1145/2047196.2047199

Shah, N and Zhou, D. (2016). No Oops, You Won't Do It Again: Mechanisms for Self-correction in Crowdsourcing. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. PMLR, New York, New York, USA, 1–10. http://proceedings.mlr.press/v48/shaha16.html

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)*. ACM, New York, NY, USA, 2645–2656. http://doi.acm.org/10.1145/2858036.2858121

Vaughan, J. W. (2018). Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46. http://jmlr.org/papers/v18/17-234.html

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.

Xia, H, Wang, Y, Huang, Y, and Shah, A. (2017). "Our Privacy Needs to Be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 113 (Dec. 2017), 22 pages. DOI: http://dx.doi.org/10.1145/3134748

Yan, Y, Rosales, R, Fung, G, and Dy, J. G. (2011). Active Learning from Crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. Omnipress, USA, 1161–1168. http://dl.acm.org/citation.cfm?id=3104482.3104628

Zhang, J, Wu, X, and Sheng, V. S. (2016). Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* 46, 4 (01 Dec 2016), 543–576. DOI:http://dx.doi.org/10.1007/s10462-016-9491-9