Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

Motivating Novice Crowd Workers through Goal Setting: An Investigation into the Effects on Complex Crowdsourcing Task Training

Amy Rechkemmer, Ming Yin
Purdue University
{arechke, mingyin}@purdue.edu

Abstract

Training workers within a task is one way of enabling novice workers, who may lack domain knowledge or experience, to work on complex crowdsourcing tasks. Based on goal setting theory in psychology, we conduct a randomized experiment to study whether and how setting different goals—including performance goal, learning goal, and behavioral goal—when training workers for a complex crowdsourcing task affects workers' learning perception, learning gain, and post-training performance. We find that setting different goals during training significantly affects workers' learning perception, but overall does not have an effect on learning gain or post-training performance. However, higher levels of learning gain can be obtained when setting learning goals for workers who are highly learning-oriented. Additionally, giving workers a challenging behavioral goal can nudge them to adopt desirable behavior meant to improve learning and performance, though the adoption of such behavior does not lead to as much improvement as when the worker decides to take part in the behavior themselves. We conclude by discussing the lessons we've learned on how to effectively utilize goals in complex crowdsourcing task training.

Introduction

Online crowdsourcing has made it easy for researchers and professionals to collect annotated data, public opinions, and other human-specific knowledge quickly and easily. Early practice of crowdsourcing often solicits human labor on relatively simple and straight-forward tasks that require only basic human skills such as visual perception (Russakovsky et al. 2015), emotion interpretation (Mohammad and Turney 2013), and relevance judgment (Alonso and Baeza-Yates 2011). More recently, the increasing demand for reaping the benefits of distributed work in complex domains leads to substantial efforts to accommodate complex tasks requiring sophisticated domain knowledge in crowdsourcing settings.

A variety of approaches have been developed to enable the completion of complex tasks by the crowd. For example, one common solution is to create a workflow to decompose complex tasks into smaller and simpler tasks that can be easily accomplished by crowd workers (Bernstein et al. 2010; Kittur et al. 2011; Little et al. 2010; Chilton et al. 2013), or potentially subcontract part of the work to other workers with relevant domain knowledge or skills (Morris et al. 2017). Sophisticated systems and algorithms have been built to organize the crowd as a dynamic team or a hierarchical organization to jointly work on the complex tasks (Retelny et al. 2014; Valentine et al. 2017; Zhou, Valentine, and Bernstein 2018). Yet another approach is to *train* workers within a complex task to prepare them with necessary knowledge and strategies for completing the task. A wide range of training methods have been studied, including training by examples and gold standard tasks (Mitra, Hutto, and Gilbert 2015; Liu et al. 2016), training by self-evaluation, expert assessment or peer feedback (Dow et al. 2012; Zhu et al. 2014; Doroudi et al. 2016), and training by communication with peers (Chen et al. 2019; Tang, Yin, and Ho 2019).

In traditional educational or organizational settings, regardless of what training method is used, an important aspect in motivating students or employees to enhance learning and performance is to *set goals*. For instance, multiple types of goals have been used for motivating individuals to learn, such as *performance goals* (i.e., goals that specify the targeted end results of learning), *learning goals* (i.e., goals that state the range of skills and knowledge the learner aims to master upon completion of the training), and *behavioral goals* (i.e., goals that describe a set of desirable behaviors the learner should follow during the learning process). Furthermore, the effects of goals on learning and performance are found to be dependent on both the type of the goal and the initiator of the goal (i.e., who sets the goal) (Latham and Brown 2006; Seijts et al. 2004; Clark et al. 2016).

The effects of setting goals when training crowd workers for complex tasks, however, are under-explored. On one hand, many positive observations on the effectiveness of goals in motivating higher levels of learning and better performance in traditional educational or organizational settings seem to suggest goal setting as a very easy-to-implement, yet promising method that can be used to improve training on any kind of complex crowdsourcing tasks. On the other hand, the nature of crowd work—that crowd workers generally have a short-term contract with requesters and may not be able to apply the domain knowledge they learn from one task to other tasks—makes it difficult to predict whether setting goals for crowd workers when training them would lead to a significant impact on them, as it does on the student or employee population.

Thus, in this paper, we seek to fill in this gap and examine the effects of goal setting on worker training in complex crowdsourcing tasks. Specifically, we design and conduct an experiment on Amazon Mechanical Turk, in which workers are recruited to perform a task that requires substantial nutritional knowledge. The experiment is divided into three phases. We provide training to workers in phase 1 to prepare workers with necessary nutritional knowledge through a combination of interactive tutorials, examples, gold standard tasks, and expert feedback. Then, we ask workers to complete real nutrition tasks during phase 2 and phase 3, which take place two days or one week after the training, respectively. In total, we consider 7 experimental treatments in our study that differ in whether and what type of goal is set (i.e., performance goal, learning goal, or behavioral goal) during the training stage, as well as who sets the goal (i.e., set by workers themselves or set by requesters).

Overall, we find that setting different goals when training crowd workers in a complex task significantly influences workers' learning perceptions, but does not affect their learning gain or post-training task performance. However, for the subset of workers who have high learning goal orientation, setting learning goals for them does lead to higher learning gain. Additionally, workers given a challenging behavioral goal perform significantly more of the desirable behavior, and workers who performed more of this behavior had higher learning gain and post-training performance, but only when they did so of their own volition. We conclude by discussing the similarity and differences of our observed effects of goal setting on training crowd workers in complex tasks as compared to those effects observed in traditional educational or organizational settings, cautioning readers to be careful in generalizing our results due to study limitations. We further reflect on the potential reasons behind the differences in these effects, and we offer a few practical lessons that we have learned from our study about how requesters can better utilize goal setting in their complex task training.

Related Work

A number of previous studies have developed a set of effective methods to train crowd workers in complex crowdsourcing tasks (Le et al. 2010; Suzuki et al. 2016; Liu et al. 2016; Zhu et al. 2014; Doroudi et al. 2016; Bigham et al. 2017; Gadiraju, Fetahu, and Kawase 2015; Gadiraju and Dietze 2017). For instance, Le et al. (2010) generalized traditional instruction methods to crowdsourcing contexts and taught workers to classify search queries by providing instructions on how to approach the task and showing detailed examples with solutions as well as the reasoning behind solutions. Suzuki et al. (2016) proposed an innovative approach of creating mentor-mentee partnerships, which enabled experienced crowd workers to help novice workers develop their skills through instruction and feedback. Comparative study (Doroudi et al. 2016) has also been conducted on the effectiveness of different training methods in crowdsourcing, and it was shown that having workers validate the work of their peers can potentially be even more helpful than having workers review expert examples.

Goal setting is an important motivating strategy that has been intensively studied in psychology. In theory, the most effective goals are both specific and difficult without being impossible (Locke and Latham 1990). In recent years, goals have been heavily studied in various educational or organizational settings to enhance learning and performance (Schunk 1990; Ames and Archer 1988; Latham and Brown 2006; Seijts et al. 2004; Clark et al. 2016). Different types of goals have been proposed, including performance goal, learning goal, and behavioral goal (Latham and Seijts 2016). It was found that students with learning goals (i.e. goals with the aim to acquire knowledge) perform better than students with specific performance goals in the end-of-semester course evaluation (Latham and Brown 2006). In addition, Clark et al. (2016) observed that workers with a behavioral goal (i.e., goal to perform a specific action that may lead to greater learning and performance) achieved better performance compared to workers with a performance goal. Other studies also showed that the effects of goals may vary with who sets the goal (Erez, Earley, and Hulin 1985; Latham, Erez, and Locke 1988). In understanding the underlying motivational processes of individuals, researchers have also identified that different people have a different "goal orientation," that is, the primary factors that motivate the individual (Dweck 1986; Bell and Kozlowski 2002), and the individuals with different types of goal orientation also respond to various types of goals differently (Seijts et al. 2004; Button, Mathieu, and Zajac 1996).

In crowdsourcing settings, the effects of goals have only been examined in the context of motivating workers to complete more tasks (Kobren et al. 2015) or implicitly incentivizing high-quality work from workers through performance-contingent financial incentives (Yin, Chen, and Sun 2014). To the best of our knowledge, this is the first study that investigates the motivating effects of goals on training workers towards better learning and performance in complex crowdsourcing tasks.

Study Design

To understand the effects of setting goals in the training stage of complex crowdsourcing tasks, we designed and conducted an experiment on Amazon Mechanical Turk (MTurk). Before running the experiment, we pre-registered our research questions, study design, and analysis methods[1], and we informally state our main research questions here:

- **Q1**: How does setting different goals affect workers' learning perceptions during training?
- **Q2**: How does setting different goals affect workers' learning gain during training and performance on tasks after training?

From the requester's point of view, understanding the effects of goals on learning outcome and post-training perfor-

[1] See https://aspredicted.org/u2pz7.pdf.

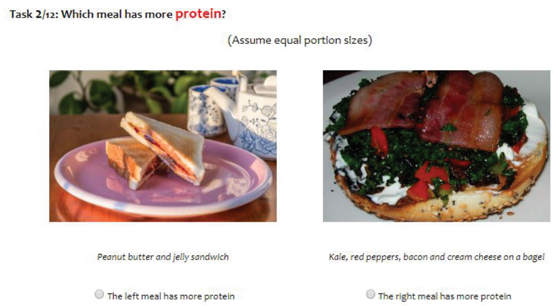


Figure 1: An example of the nutrition task.

mance (i.e., Q2) provides direct implications for how to design effective goals in their crowdsourcing tasks. We also included examination of the effects of goals on workers' learning perceptions (i.e., Q1) as learning perceptions may reflect worker satisfaction, which can potentially further influence the amount of effort workers put into the tasks and worker retention. In addition, we also pre-registered our intent to conduct exploratory analysis to examine when and why different goals work or do not work in influencing workers' learning and performance.

Experimental Tasks

In this experiment, we used a task that asked workers to identify nutritional components in meals. In particular, in each task, a worker was given two photographs of meals along with descriptions of the main ingredients in each meal. The worker was then asked to identify which of the two meals contained more of a specified nutritional component. Four nutritional components—fat, fiber, protein, and carbohydrates—were examined in these tasks. Figure 1 shows an example of the nutrition task. Photographs used in these tasks were taken from Burgermaster et al. (2017).

We chose this task for our experiment as it requires substantial nutritional knowledge, which is not a common skill among laypeople. Indeed, the task of nutrition analysis has been previously used as a complex task by a number of crowdsourcing researchers to show how to facilitate learning (Burgermaster et al. 2017) and how to design workflows to decompose complex tasks (Noronha et al. 2011).

Experimental Procedure

Our experiment was divided into three phases. Phase 1 was the "training" phase and was used to prepare workers with necessary nutritional knowledge before they complete actual nutrition tasks. Phases 2 and 3 were the work phases in which workers could use the knowledge they learned from phase 1 to complete a sequence of nutrition tasks. Separate pools of tasks were created to be sampled from for each of the phases, and through a pilot study we found no differences in difficulty for tasks in different pools.

Phase 1 Figure 2 displays the overall flow of phase 1. In phase 1, a worker started by completing a session of 12 randomly-sampled nutrition tasks (3 tasks for each of the 4 nutritional components), which we refer to as the *pre-test*. Upon completion of these tasks, we told workers their overall accuracy in these 12 tasks as well as their accuracy within each nutritional component. Next, prior to taking the nutrition lessons, the worker might be given a goal or be asked to set a goal for herself for the nutrition lessons depending on the treatment the worker was assigned (see more details in the "Experimental Treatments" subsection). The worker was asked to keep this goal in mind throughout the lessons.

The worker was then required to go through 4 nutrition lessons, each corresponding to one nutritional component. The lesson for each nutritional component contained both textual information and one short video teaching workers what the component is and what foods are rich in it. At the end of each lesson, there was a qualification question checking whether the worker understood the information in the lesson. The worker could take the 4 lessons in any order that she wanted, but to proceed on to the next section, she had to answer the qualification questions for all 4 lessons correctly.

After the nutrition lessons, the worker was given an option of taking up to 10 practice tasks. The practice tasks were in the same form as the nutrition tasks. For each practice task that the worker took, we provided her feedback on both the correct answer and an explanation. Similar to that in Clark et al. (2016), workers were *not* required to take these practice tasks in our experiment, which allowed us to observe how different goals may affect workers' learning and performance in the tasks through influencing their tendency to adopt desirable behavior (e.g., take more practice tasks).

At the end of phase 1, the worker completed the following steps. First, she took a questionnaire on her goal orientation. We adopted the scales from Button, Mathieu, and Zajac (1996) to measure the worker's goal orientation in terms of performance (i.e., the level of motivation towards achieving high performance) or learning (i.e., the level of motivation towards learning new things). Second, she answered two survey questions on a 5-point scale regarding her perception of learning in the nutrition lessons:

- *Helpfulness*: How helpful did you find the nutrition lessons?
- *Learning*: How much do you feel you have learned from the nutrition lessons you went through earlier?

Third, the worker completed another session of 12 randomly-sampled *post-test* nutrition tasks (again 3 tasks for each nutritional component), reviewed her performance in them, and evaluated whether she had achieved her goal (if applicable). Finally, the worker reported basic demographic information, including age, gender, geographical location, and prior nutritional knowledge, through an exit survey.

Phases 2 and 3 Phase 2 was conducted two days after phase 1 took place, and phase 3 happened one week after phase 1. Regardless of the worker's treatment, in both phases 2 and 3, she was asked to complete a random sequence of 12 nutrition tasks that she had not seen before, again, with 3 tasks for each nutritional component. We did not provide any feedback on answer accuracy for tasks in phases 2 and 3. Thus, the worker's accuracy in phases 2 and 3 reflected

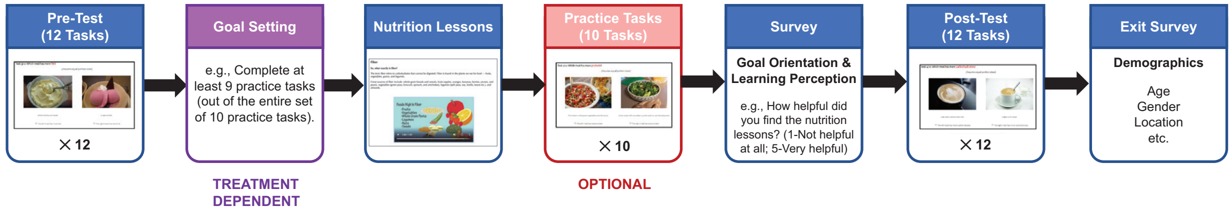


Figure 2: A diagram of sections that workers went through in phase 1 of our experiment.

her performance on real-world tasks, either shortly after the training or a while after receiving the training.

Experimental Treatments

To examine the effects of goal setting, we varied whether and how goals were set for workers before they took the nutrition lessons in phase 1. We considered a 2×3 design along two factors: the *initiator* of the goal and the *type* of the goal.

More specifically, the initiator of the goal can be either the worker herself or the requester. That is, when the initiator of the goal was the worker, she was asked to set a goal for herself before taking the nutrition lessons. However, when the initiator of the goal was the requester, the worker was given a goal that was pre-determined by us. In addition, we considered three types of goals in our experiment:

- *Performance goal*: This goal specifies the number of *post-test* tasks that the worker should answer *correctly*. When a worker received a performance goal from the requester, the goal was to answer at least 10 out of 12 possible post-test questions correctly. When a worker was asked to set a performance goal herself, she could choose any integer number between 0 and 12 of post-test tasks to aim to answer correctly.
- *Learning goal*: This goal specifies the kind of knowledge the worker aims to learn from the nutrition lessons. For workers who were given a learning goal, the goal was stated as "Learn and recognize the types of foods that are high in carbohydrates, protein, fiber, and fat." On the other hand, workers who were asked to set a learning goal for themselves used free-form language to create their own goals in terms of the main concepts or ideas that they wanted to learn from the nutrition lessons.
- *Practice goal*: This goal is an operationalization of the "behavioral goal" and specifies the number of practice tasks that the worker should complete after taking the nutrition lesson. When a practice goal was assigned by the requester, the goal was to complete at least 9 out of 10 possible practice tasks. When a worker set a practice goal for herself, she could choose any integer number of practice tasks between 0 and 10 to aim to complete.

As previous research suggests challenging goals are more motivating (Locke and Latham 1990), we intentionally designed the request-set goals to be difficult to achieve. In contrast, for goals that were set by workers, it was entirely up to workers themselves to determine whether their goals were easy or challenging. To remind workers of the goals that were either assigned to them or set by themselves, throughout the nutrition lessons, we displayed their goal at the top of the webpages which contained the training material.

Finally, we also included a control treatment where the worker was not given a goal nor asked to set a goal for herself. Together with the previous 6 treatments, in total, we had 7 treatments in this experiment.

Other Experimental Control

Our experiment was implemented as a Human Intelligence Task (HIT) and was open only to U.S. workers who had completed at least 500 HITs on MTurk previously. For phase 1, each worker was *randomly* assigned to one of the 7 treatments upon arrival and received a fixed payment of \$2 after they submitted the HIT. We then invited all workers who had submitted the phase 1 HIT (and only these workers) back to participate in phases 2 and 3, in which they received a fixed payment of \$0.75[2]. Each worker was allowed to take the HIT for each phase at most once.

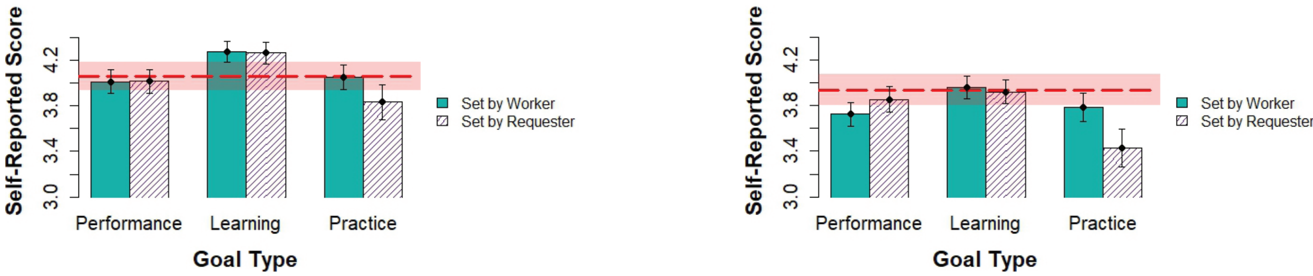
Data

In total, 659 workers participated in our experiment in phase 1 [3]. Among these workers, 58.3% of them were female, and their average age was 38.2. When asked to report their prior nutritional knowledge level from 1 ("No prior knowledge relating to nutrition") to 5 ("A lot of prior knowledge"), workers' average reported level was 2.98. Across all workers, we find that the mean and standard deviation for the number of pre-test questions a worker correctly answered in phase 1 was 8.5 (i.e., 71% accuracy) and 1.6. About 11% of the workers answered no more than 6 pre-test questions correctly, meaning their performance in the nutrition task prior training was no better than random, which again confirmed the difficult nature of this task for laypeople. We observed no significant difference in demographics, prior knowledge level, or the number of pre-test questions correctly answered across workers of different treatments in phase 1.

In phase 1, on average, workers who were asked to set their own performance goals aimed to answer at least 8.89

[2]The effective hourly wage of the phase 1 HIT was \$8/hour, and \$15/hour for phase 2 and phase 3 HITs.

[3]We conducted a pilot study and used power analysis to determine the sample size. Based on an insignificant effect of goals on learning gain, we found the sample size needed was at least 588 workers given a power of 0.9 and an alpha of 0.05.



(a) Perceived helpfulness of the lessons

(b) Perceived learning level in the lessons

Figure 3: Workers' perceptions of learning across different treatments. The mean value of self-reported scores for each treatment is plotted, and error bars represent the standard errors of the mean. For the control treatment, the dashed horizontal lines represent the mean values, and the standard errors of the mean are shown by the red shaded areas.

out of 12 post-test questions correctly, and workers who were asked to set their own practice goals aimed to complete at least 4.3 out of 10 possible practice tasks. Workers who were asked to set their own learning goals also set meaningful learning goals such as "I want to learn more about fat content in foods" and "I'd like to learn which foods likely don't have gluten but also have high fiber." We also observed no significant difference in the amount of time a worker spent on the nutrition lessons or the number of times the worker interacted with the training material, such as playing videos in the lessons, across workers of different treatments.

Among all workers who participated in phase 1 of our experiment, 581 workers took the phase 2 experiment and 558 workers took the phase 3 experiment. We observed *no* significant demographic or prior knowledge differences between all workers who participated in our phase 1 HIT, and the subset who came back to participate in our phase 2 HIT (or phase 3 HIT). We also did *not* find any evidence suggesting that the worker's post-test performance in phase 1 changed the worker's likelihood of taking the phase 2 or phase 3 HIT.

Results

We first analyze the experimental data to understand, overall, how setting goals during the training stage of a complex crowdsourcing task affects workers' learning perception (Q1), learning gain, and post-training performance (Q2). Then, we conduct additional exploratory analyses to delve deeper into understanding why and under which conditions certain goals are effective/not effective.

Q1: The Impact on Learning Perception

We start by examining how setting goals affects workers' learning perceptions in the training stage of complex crowdsourcing tasks. We measure workers' learning perceptions using their self-reported scores on the helpfulness of the training material and the amount they have learned in the nutrition lessons, and Figures 3a and 3b show the comparison on these two metrics, respectively. Visually, it seems that setting different goals in the training stage for the nutrition task indeed has some impact on workers' perceptions of learning, and a learning goal seems to lead to the highest level of learning perception. To validate our visual intuition, we conducted statistical tests to examine whether the distributions for workers' learning perceptions in different treatments are statistically the same. First, a one-way ANOVA test across all 7 treatments suggests a marginally significant difference in workers' scores on the helpfulness of nutrition lessons ($p = 0.096$) and workers' perceived levels of learning ($p = 0.075$). Thus, setting different goals indeed affects workers' learning perceptions, though post-hoc Tukey HSD tests suggest no significant differences on learning perceptions between any of the treatments with some goal and the control treatment with no goal.

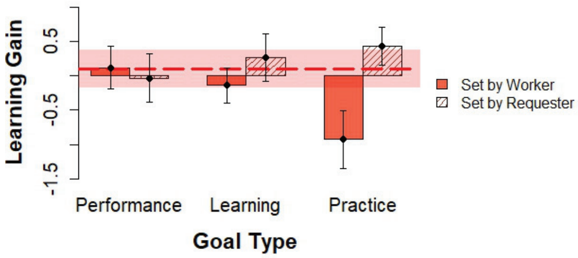
To further understand how the initiator and type of the goal affect workers' perceptions of learning, we conducted a two-way ANOVA on the data obtained from all but the control treatment. Doing so, we found that the type of the goal has a significant effect on both the reported helpfulness of the lessons ($p = 0.008$) and the perceived level of learning ($p = 0.036$). Post-hoc Tukey HSD test results further suggest that workers with a learning goal reported the nutrition lessons as significantly more helpful than workers with a performance goal ($p = 0.037$) or workers with a practice goal ($p = 0.015$), and they also perceived themselves as having learned significantly more from the nutrition lessons compared to workers with a practice goal ($p = 0.028$).

On the contrary, we found the initiator of the goal does not have a significant effect on either metric of learning perception ($p = 0.509$ for the helpfulness question and $p = 0.552$ for the learning question). We did not detect any significant interactions between the type of goals and the initiator of goals on workers' learning perceptions, either.

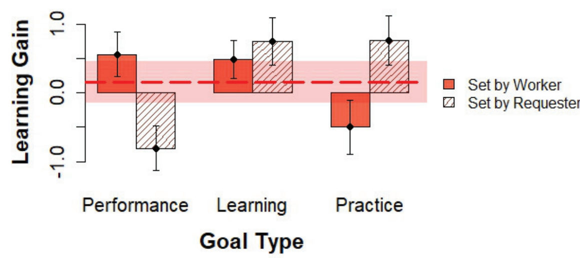
Q2: The Impact on Learning Gain and Post-Training Performance

We now move on to understand whether setting goals can effectively lead to different levels of learning during training and different levels of performance on tasks after training.

To quantify how much a worker actually learned in the nutrition lessons, we define a worker's learning gain as the dif-



(a) Workers with high performance goal orientation



(b) Workers with high learning goal orientation

Figure 4: The learning gain across different treatments for workers with high performance/learning goal orientation. The mean value of the learning gain for each treatment is plotted, and error bars represent the standard errors of the mean. Dashed lines and red shaded areas show the mean values and standard errors of the mean for learning gain in the control treatment.

ference in the number of pre-test and post-test questions that she answered correctly in phase 1. A one-way ANOVA test across all 7 treatments suggests no statistically significant difference in learning gain across different treatments ($p = 0.656$). Similar as before, when we dropped the data from the control treatment and conducted a two-way ANOVA test on the other 6 treatments, we still found that neither the initiator of the goal nor the type of the goal has any significant impact on how much the worker actually learned during the training ($p = 0.255$ for goal initiator and $p = 0.964$ for goal type), and there is no significant interaction effect either.

Next, we analyzed the experimental data that we collected from phases 2 and 3 of our experiment to understand whether setting different goals during the training stage leads to significant differences in post-training performance. We again conducted one-way ANOVA tests for the performance data across all 7 treatments collected in phase 2 and phase 3, separately. No significant difference was found for workers' accuracy in the nutrition tasks in phase 2 ($p = 0.787$) or phase 3 ($p = 0.713$) across treatments. We further conducted two-way ANOVA tests on all but the control treatment to investigate the effects of goal initiator and goal type on post-training performance, and again, no significant effects were detected (i.e., phase 2: goal type $p = 0.886$, goal initiator $p = 0.391$; phase 3: goal type $p = 0.823$, goal initiator $p = 0.159$).

In other words, with respect to all the workers who took our phase 1 HIT, setting different goals when training them for the nutrition task does *not* lead to significantly different learning outcomes or post-training task performance.

Exploratory Analysis

So far, we have learned that for our full population of workers, setting different goals during the training stage of complex crowdsourcing tasks affects workers' learning perceptions, but has no obvious effect on the actual learning gain during training or post-training performance on real-world tasks. This is in contrast with the effects of goals observed in traditional educational or organizational settings, which motivated us to look deeper into when and why various goals may work or not work in the crowdsourcing context.

On the one hand, there is reason to believe that some goal types may be more effective in influencing learning and performance for certain subsets of workers. For example, previous research suggests that individuals have different types of goal orientation (Dweck 1986; Bell and Kozlowski 2002) and may respond to various goals differently depending on whether or not the goal *matches* with their goal orientation. On the other hand, goals like the behavioral goal are designed to motivate people through well-understood mechanisms (e.g., encourage the adoption of desirable behavior). Since we did not see that setting behavioral goals for workers leads to any significant improvement in learning gain or performance, we seek to explore the reason why here.

We are therefore interested in, and have pre-registered our intent to explore, the following additional questions:

- Does setting a goal for workers that matches with their goal orientation lead to higher levels of learning gain and post-training performance?
- Does setting a practice goal lead to more practice tasks being completed, and does completing more practice tasks associate with higher levels of learning gain and post-training performance?

The Role of Goal Orientation on the Effectiveness of Goals We first explore how workers' goal orientation moderates the effects of goals. We used a median split to classify each worker as "high" or "low" on performance (or learning) goal orientation based on her responses to the goal orientation scales during phase 1. Such classification enabled us to look into the effects of different goals on the subset of workers who have high performance goal orientation and the subset of workers who have high learning goal orientation separately.

With respect to workers who are substantially motivated to obtain high performance in tasks, Figure 4a displays how setting different goals in the training stage affects their learning gain. For them, we found no significant difference in learning gain across the 7 treatments through one-way ANOVA ($p = 0.156$), and the two-way ANOVA test on data from all but the control treatment indicates the effect of goal type on learning gain is also not significant ($p = 0.521$). This means that setting a performance goal for workers who have high performance goal orientation does *not* lead to significantly higher levels of learning gain compared to the cases when no goal or other types of goals are set. Moreover, we did not find any significant differences on post-training performance across workers with high performance goal orientation who were assigned to different treatments (one-way ANOVA: $p = 0.642$ for phase 2 and $p = 0.998$ for phase 3).

On the other hand, Figure 4b shows the impact of goals on learning gain when restricted to workers who are more motivated to learn new things. Here, it seems that setting learning goals for these workers with high learning goal orientation consistently implies a relatively high level of learning outcome. Results of a one-way ANOVA test also suggest that there is a statistically significant difference in learning gain across all 7 treatments ($p = 0.004$).

To further see how the type and initiator of the goal affects learning gain for workers with high learning goal orientation, we conducted a two-way ANOVA with all but the control treatment, and results show that while the initiator of the goal does not significantly affect the learning gain ($p = 0.851$), the type of goal does ($p = 0.049$). In particular, post-hoc Tukey HSD tests revealed that for workers who are highly motivated to learn new things, a learning goal leads to higher levels of learning gain compared to a performance goal when the goal is set by the requester ($p = 0.025$). We also detected a significant interaction effect between goal type and goal initiator on learning gain ($p = 0.002$). As shown in Figure 4b, for workers highly oriented for learning, giving them a performance goal results in a worse learning outcome compared to a performance goal set by the worker herself, while having the worker setting a practice goal herself results in a worse learning outcome than giving a practice goal to the worker. The higher levels of learning gain brought up by learning goals, however, do not translate into higher post-training performance—among workers with high learning goal orientation, the phase 2 and phase 3 performance across all 7 treatments is still statistically the same (one-way ANOVA: $p = 0.881$ for phase 2 and $p = 0.926$ for phase 3), and the initiator or type of the goal casts no significant impact on post-training performance.

In sum, we found that matching a performance goal to workers who are highly oriented for performance does not improve either the learning gain or post-training performance, while matching a learning goal to workers who are highly oriented for learning improves the learning gain, but does not affect the post-training performance.

Why Behavioral Goals Don't Work? Among the three types of goals that we have experimented with in this study, the behavioral goal (which is operationalized as the practice goal) seems to not only exhibit limited impact on workers' learning gain or post-training performance, but also leads to the lowest level of learning perceptions (see Figure 3). Naturally, one may wonder why behavioral goals seem to be ineffective in complex crowdsourcing task training.

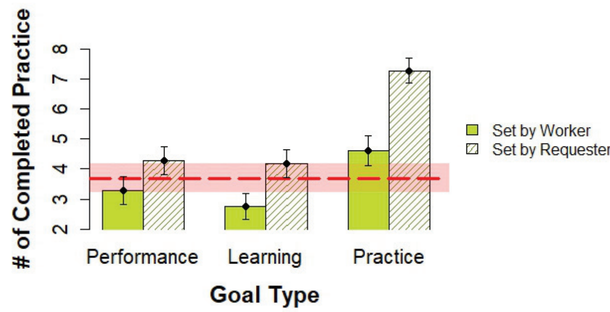


Figure 5: The number of practice tasks completed by workers across different treatments. The mean value for each treatment is plotted, and error bars represent the standard errors of the mean. The dashed line and red shaded area represent the mean value and the standard error of the mean of completed practice tasks in the control treatment.

First, we note that when workers were asked to set a practice goal for themselves, they tended to set an "easy" goal—on average, they aimed at completing 4.3 practice tasks, while workers who were given a practice goal were told to try to complete at least 9 practice tasks. Figure 5 shows the number of practice tasks that workers of different treatments actually completed, in which we observed a statistically significant difference ($p < 0.001$). In particular, post-hoc Tukey HSD tests show that workers who were given a practice goal completed significantly more practice tasks than workers in all other treatments ($p < 0.005$). On the other hand, we detected no significant difference at $p < 0.05$ level between the number of practice tasks completed by workers who set their own practice goals and workers with no goal or had performance or learning goals.

To fully understand the effectiveness of behavioral goals, we next ask whether completing more practice tasks actually associates with higher learning gain and better post-training performance for workers. Since most of the workers completed either 0 or 10 practice tasks, we split workers into two groups—the group who completed at least 5 practice tasks, and the group who completed fewer than 5 practice tasks. Conducting two-sample t-tests between these two groups of workers, we confirmed that workers who completed more practice tasks not only learned more during the training stage, but also achieved better performance in both phase 2 and phase 3 ($p < 0.001$ for all three comparisons). This result is puzzling considering that we did not see that workers who were given a practice goal—who indeed completed more practice tasks—obtained significantly higher levels of learning gain or post-training performance.

An in-depth analysis of the data suggests one possible explanation for why we see this discrepancy—while workers who were given a challenging practice goal indeed completed more practice tasks, the increase in learning gain (or post-training performance) they obtained from the extra practice is *less* than the increase that workers from other conditions experienced. For example, when considering only those workers who completed all 10 practice tasks in each treatment, workers who were given a practice goal achieved an average learning gain of 1.01, and answered 8.85 and 9.10 questions correctly in phase 2 and phase 3, respectively. In contrast, workers whose goal was not a practice goal or did not have any goal, on average, achieved a higher level of learning gain (i.e., 1.22), and completed more tasks correctly in phases 2 and 3 (i.e., 9.34 correct answers for phase 2 and 9.35 correct answers for phase 3)[4]. As in Figure 3, we observed workers who are given a practice goal seem to have the lowest level of perception of learning. We conjecture that many of our workers assigned a challenging practice goal chose to satisfice, leading to them meeting the goal even if they did not reap the benefits that those who completed the practice of their own volition did.

Conclusions and Discussions

In this paper, we investigate the effects of goal setting on training crowd workers to complete complex tasks. Overall, we find that setting different goals indeed changes worker's learning perceptions, but it has no significant impact on workers' learning gain or post-training performance. Through exploratory analysis, we find that workers with a high learning goal orientation have a significantly larger learning gain when they have a learning goal than other goal types. Additionally, we find that workers given a practice goal completed significantly more practice tasks than workers in other conditions, though they did not see the same amount of increase in learning gain and post-test performance by completing more practice tasks as the workers in other conditions who completed more practice tasks of their own volition did.

We now reflect on potential reasons for the discrepancies between our results and those from traditional goal setting studies, the takeaways from our results that can best inspire future crowdsourcing design for complex tasks, and the limitations of our study.

Differences Between Crowdsourcing and Typical Goal Setting Environments. Compared to experimental results obtained in goal setting literature in psychology, while we did get some consistent results when applying goal setting to complex crowdsourcing task training (e.g., the effects of goals vary between workers with different types of goal orientation), there are many notable discrepancies. Unlike in previous literature, we did not find support that setting different goals leads to any significant difference in post-training performance, and we found minimal evidence that goal setting had a impact on learning gain. We now address the differences between the crowdsourcing environment and traditional goal setting environments in order to conjecture about what might have caused us to see different results.

Much of the previous goal setting studies have taken place in the context of a classroom or workplace. This not only provides a relatively long time-frame for goals to be impacting learning and performance, but it is an in-person environment where students and teachers (or employees and managers) are able to interact with one another, and students/employees are held accountable for their work. In contrast, online microtask crowdsourcing environments typically lack these characteristics. The anonymity between workers and requesters and the ability to switch to a different task makes it less appealing for workers to make any unnecessary commitments. Even if workers are willing to commit to a goal, many crowdsourcing tasks may supply too short a time-frame for goals to take effect.

Another critical difference between traditional goal setting environments and ours is that in the former case, both the subject of the goal (e.g., learning specific knowledge) and the final performance outcome (e.g., final grade) are of real importance to those who set the goals. As such, people are incentivized to take their goals seriously. This is not necessarily true for workers in our study—we intentionally set the financial payments received to be independent of their performance so as to examine the motivating power of goals alone, especially since goal setting theory warns against the use of incentives that may distract from the goal (Locke 1996). Without an external incentive, such as increased payment or the promise of additional work, some workers may not have had the motivation to commit to their goal.

Practical Implications for Setting Goals When Training Crowd Workers. One of the lessons that we learned through this study is that when training workers for complex crowdsourcing tasks, simply recruiting a random set of workers and setting some goal for them may not be effective. Instead, the right type of goal needs to be set for the right kind of workers in the right way. For instance, the observations that workers with different types of goal orientation respond to various goals differently suggest the potential for *personalizing* goals for workers with different characteristics. Additional studies should be done to further explore ways to personalize goals for workers in order to see greater benefits when training for complex tasks.

Another lesson is the need to improve the design of behavioral goals in crowdsourcing settings, as the practice goals in our experiment were not very effective in influencing learning and performance, but we were able to see the potential of adopting desirable behavior. One possible way is to communicate the benefits of the desired behavior more explicitly. Take the goal of completing more practice tasks as an example. If the potential benefits of practice tasks can be better communicated to workers, then those who set their own practice goal may be willing to set a more challenging goal, and those who are given a practice goal may put greater effort into using the practice tasks to help improve their knowledge. In that case, we might be able to find the behavioral goal to be effective in influencing learning and performance.

[4]Note that for those workers who were asked to set a practice goal for themselves and chose to complete all 10 tasks, we did not see the increase in learning gain or post-training performance lessened—their average learning gain is 1.35, and the average number of correctly answered questions in phases 2 and 3 are 9.15 and 9.36, respectively.

Our work shows that behavioral goals increase adoption of desired behavior, and future work should be done to discover how to make workers see the merit of this desired behavior.

Limitations and Future Work. Our study was conducted on a single task which may be representative of tasks where substantial domain knowledge is needed to perform the task well. However, other types of complex tasks exist, such as tasks that require sophisticated problem-solving strategy and creativity. We also only consider one type of training method within our one-week study. Caution should be used when generalizing our results to other types of complex tasks, training methods or settings over a long-term period, and additional research is needed to obtain a more comprehensive understanding of the effects of goals in training workers.

A direction of future work is to explore the design space of goals and examine their effectiveness along other key design dimensions, such as the connection between goal attainment with different incentives. Since external incentive is one of the discrepancies between our study and traditional goal setting studies, we believe associating incentives that directly benefit crowd workers with the attainment of goals would be promising in improving the effects of goals. Exploring the use of goal setting on different online crowdsourcing platforms that have different inherent incentives (e.g., citizen science platforms) is an additional direction to pursue.

References

Alonso, O., and Baeza-Yates, R. 2011. Design and implementation of relevance assessments using crowdsourcing. In Clough, P.; Foley, C.; Gurrin, C.; Jones, G. J. F.; Kraaij, W.; Lee, H.; and Mudoch, V., eds., *Advances in Information Retrieval*, 153–164. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ames, C., and Archer, J. 1988. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of educational psychology* 80(3):260.

Bell, B. S., and Kozlowski, W. 2002. Goal orientation and ability: Interactive effects on self-efficacy, performance, and knowledge. *Journal of Applied Psychology* 87(3):497.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Bigham, J. P.; Williams, K.; Banerjee, N.; and Zimmerman, J. 2017. Scopist: building a skill ladder into crowd transcription. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, 2. ACM.

Burgermaster, M.; Gajos, K. Z.; Davidson, P.; and Mamykina, L. 2017. The role of explanations in casual observational learning about nutrition. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4097–4145. ACM.

Button, S. B.; Mathieu, J. E.; and Zajac, D. M. 1996. Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational behavior and human decision processes* 67(1):26–48.

Chen, Q.; Bragg, J.; Chilton, L. B.; and Weld, D. S. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Chilton, L. B.; Little, G.; Edge, D.; Weld, D. S.; and Landay, J. A. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 1999–2008. New York, NY, USA: ACM.

Clark, D.; Gill, D.; Prowse, V.; and Rush, M. 2016. Using goals to motivate college students: Theory and evidence from field experiments. *Review of Economics and Statistics* 1–45.

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*, CHI '16, 2623–2634. New York, NY, USA: 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*, CSCW '12, 1013–1022. New York, NY, USA: ACM.

Dweck, C. S. 1986. Motivational processes affecting learning. *American psychologist* 41(10):1040.

Erez, M.; Earley, P. C.; and Hulin, C. L. 1985. The impact of participation on goal acceptance and performance: A two-step model. *Academy of Management journal* 28(1):50–66.

Gadiraju, U., and Dietze, S. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 105–114.

Gadiraju, U.; Fetahu, B.; and Kawase, R. 2015. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World*. Springer. 100–114.

Kittur, A.; Smus, B.; Khamkar, S.; and Kraut, R. E. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 43–52. ACM.

Kobren, A.; Tan, C. H.; Ipeirotis, P.; and Gabrilovich, E. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th international conference on world wide web*, 592–602. International World Wide Web Conferences Steering Committee.

Latham, G. P., and Brown, T. C. 2006. The effect of learning vs. outcome goals on self-efficacy, satisfaction and performance in an mba program. *Applied Psychology* 55(4):606–623.

Latham, G. P., and Seijts, G. H. 2016. Distinguished scholar invited essay: Similarities and differences among performance, behavioral, and learning goals. *Journal of Leadership & Organizational Studies* 23(3):225–233.

Latham, G.; Erez, M.; and Locke, E. 1988. Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the erez-latham dispute regarding participation in goal setting. *Journal of Applied Psychology* 73:753–772.

Le, J.; Edmonds, A.; Hester, V.; and Biewald, L. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *In SIGIR 2010 workshop*, 21–26.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2010. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 57–66. ACM.

Liu, A.; Soderland, S.; Bragg, J.; Lin, C. H.; Ling, X.; and Weld, D. S. 2016. Effective crowd annotation for relation extraction. In *HLT-NAACL*.

Locke, E. A., and Latham, G. P. 1990. *A theory of goal setting & task performance.* Prentice-Hall, Inc.

Locke, E. A. 1996. Motivation through conscious goal setting. *Applied and preventive psychology* 5(2):117–124.

Mitra, T.; Hutto, C.; and Gilbert, E. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 1345–1354. New York, NY, USA: ACM.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Morris, M. R.; Bigham, J. P.; Brewer, R.; Bragg, J.; Kulkarni, A.; Li, J.; and Savage, S. 2017. Subcontracting microwork. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 1867–1876. ACM.

Noronha, J.; Hysen, E.; Zhang, H.; and Gajos, K. Z. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 1–12. ACM.

Retelny, D.; Robaszkiewicz, S.; To, A.; Lasecki, W. S.; Patel, J.; Rahmati, N.; Doshi, T.; Valentine, M.; and Bernstein, M. S. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 75–85. ACM.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Schunk, D. H. 1990. Goal setting and self-efficacy during self-regulated learning. *Educational psychologist* 25(1):71–86.

Seijts, G. H.; Latham, G. P.; Tasa, K.; and Latham, B. W. 2004. Goal setting and goal orientation: An integration of two different yet related literatures. *Academy of management journal* 47(2):227–239.

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*, 2645–2656. ACM.

Tang, W.; Yin, M.; and Ho, C.-J. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*, WWW '19, 1794–1805. New York, NY, USA: ACM.

Valentine, M. A.; Retelny, D.; To, A.; Rahmati, N.; Doshi, T.; and Bernstein, M. S. 2017. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3523–3537. New York, NY, USA: ACM.

Yin, M.; Chen, Y.; and Sun, Y.-A. 2014. Monetary interventions in crowdsourcing task switching. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Zhou, S.; Valentine, M.; and Bernstein, M. S. 2018. In search of the dream team: Temporally constrained multi-armed bandits for identifying effective team structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 108:1–108:13. New York, NY, USA: ACM.

Zhu, H.; Dow, S. P.; Kraut, R. E.; and Kittur, A. 2014. Reviewing versus doing: learning and performance in crowd assessment. *Conference on Computer Supported Cooperative Work* 1445–1455.