# Learning from the Crowd: Observational Learning in Crowdsourcing Communities

**Lena Mamykina**
Columbia University
New York, NY, USA
lena.mamykina@dbmi.columbia.edu

**Thomas N. Smyth, Jill P. Dimond**
Sassafras Tech Collective
Ann Arbor, MI, USA
(jill, tom}@sassafras.coop

**Krzysztof Z. Gajos**
Harvard SEAS
Cambridge, MA, USA
kgajos@eecs.harvard.edu

## ABSTRACT
Crowd work provides solutions to complex problems effectively, efficiently, and at low cost. Previous research showed that feedback, particularly correctness feedback can help crowd workers improve their performance; yet such feedback, particularly when generated by experts, is costly and difficult to scale. In our research we investigate approaches to facilitating continuous observational learning in crowdsourcing communities. In a study conducted with workers on Amazon Mechanical Turk, we asked workers to complete a set of tasks identifying nutritional composition of different meals. We examined workers' accuracy gains after being exposed to expert-generated feedback and to two types of peer-generated feedback: direct accuracy assessment with explanations of errors, and a comparison with solutions generated by other workers. The study further confirmed that expert-generated feedback is a powerful mechanism for facilitating learning and leads to significant gains in accuracy. However, the study also showed that comparing one's own solutions with a variety of solutions suggested by others and their comparative frequencies leads to significant gains in accuracy. This solution is particularly attractive because of its low cost, minimal impact on time and cost of job completion, and high potential for adoption by a variety of crowdsourcing platforms.

### Author Keywords
Crowdsourcing, nutritional assessment, observational learning.

### ACM Classification Keywords
H5.3. Group and Organization Interfaces: Web-based Interaction

## INTRODUCTION
In recent years, crowd computing emerged as a powerful alternative to strictly computational approaches to solving a variety of problems [2], [1], [7], [18]. The benefits of crowd computing are beyond doubt: it provides solutions to complex problems effectively, efficiently, and at low cost. Crowd computing is particularly effective for completing tasks that require human perception, judgment and common sense. Such tasks are frequently beyond the reach of computers, yet they can be solved with little effort by people. Crowd computing is less commonly used for tasks that require special knowledge and skills, such as visual design, coding and programming, and nutritional assessment of meals. Tasks like these typically require both domain and discipline-specific knowledge, as well as awareness of social norms, practices, and conventions related to these disciplines. One approach to enabling crowdsourcing for these tasks is through expert-based communities, such as 99design.com that focuses on graphic design. However, these specialized communities might present high entry barriers for crowd workers. An attractive alternative to searching for existing expertise is to develop mechanisms for training crowd workers on the job and helping them acquire the necessary knowledge and skills. This approach would benefit the requesters, who could receive higher quality solutions. In addition, it would benefit the workers and allow them to acquire and develop new skills, grow expertise and, potentially, advance their careers [14].

For crowdsourcing tasks that rely on general human abilities and common sense (such as writing product reviews), recent research has demonstrated that self-assessment, assessing the work of others, and expert feedback can all result in improved performance over time [11],[23]. Less is known, however, about how to improve crowd workers' performance on more specialized tasks discussed above. A common approach to promoting learning for such tasks in traditional learning environments is through explicit instruction coupled with individualized correctness feedback on practice problems, typically generated by experts and accompanied by explanations of errors [23],[11]. Yet in a crowdsourcing environment, neither of these may be readily available or feasible. Explicit instruction may require time investment from both job requesters and workers. Moreover, for the vast majority of crowdsourcing jobs, the correct or expert-generated solutions do not exist.

As an alternative to relying on experts, we investigate the effectiveness of peers as a source of feedback for improving performance on knowledge-based crowd computing


*CHI'16*, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858560

tasks. Typically, such feedback is generated when peers explicitly assess accuracy of an individual's solutions. There is substantial prior research showing that in traditional learning environments feedback generated by peers can be easier to understand and integrate in one's own work than teacher-generated feedback, and that critical feedback from peers is less demoralizing [19], [10]. On the other hand, peer feedback may not always be correct. Moreover, generating such feedback in crowdsourcing environments may increase both the time required to complete the job, and its cost. An alternative, more implicit approach to generating peer feedback in crowd communities is by helping workers to compare their own solutions with solutions generated by the crowd (Figure 1). This type of feedback is particularly attractive because it does not impact either time or cost of job completion and takes advantage of people's natural tendency for *observational learning* [6].

To investigate the effectiveness of peer feedback and observational learning and to compare it with the more traditional expert-generated feedback, we conducted a study with workers on Amazon Mechanical Turk (MTurk). In this study, workers were asked to complete a set of tasks and either received no feedback, received expert-generated feedback on correctness, considered here as gold standard, or received one of the two types of peer-generated feedback discussed above, explicit or implicit. Across conditions, we relied on active learning style [9] in which individuals can not only receive feedback, but also immediate apply this feedback to their solutions.

The domain for the study was nutritional assessment of meals; the specific task required workers to match ingredients of photographed meals to different macronutrients, including protein, fat, carbohydrate, and fiber, while looking at photographs of these meals. A professional dietitian developed gold standard for all tasks in the study; individuals' answers were compared with gold standard, thus establishing accuracy of their solutions.

In this study, our main focus was on accuracy gains for individual workers, and the difference in these gains for different types of feedback mechanisms. Our specific research questions included the following:

1. Can peer-generated feedback have a positive impact on workers' objective performance and learning gains, as well as their self-efficacy and perceived learning as compared to competing tasks without feedback or as compared to expert-generated feedback?
2. What form of peer-generated feedback (explicit or implicit) leads to the best performance and highest learning gains?

***The results of the study suggested that while expert-generated feedback led to the highest improvement in accuracy of workers' solutions, implicit peer-generated feedback through comparison between workers' own solutions to solutions generated by the crowd also led to significant gains in workers' accuracy, albeit with a smaller effect size***. Of these two solutions, the latter one has a particular advantage because it does not rely on involvement of experts, does not increase workload of individual workers and thus has minimal negative consequences on time required for job completion and its cost, is highly scalable and can potentially be utilized across a variety of crowdsourcing jobs. In contrast, explicit peer feedback did not result in any improvement in workers' accuracy.

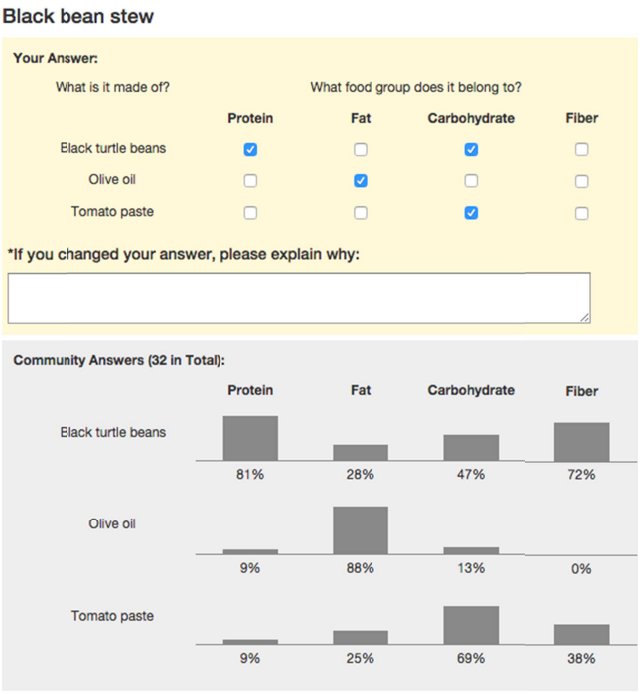


**Figure 1: Design of the implicit peer feedback showing comparison of a worker's own solution with distribution of solutions submitted by other Turkers.**

## THEORETICAL FOUNDATIONS

### Observational Learning

The solutions explored in this paper draw on social theories of learning and, specifically, on *observational learning*. Observational learning has been explored within several theories of learning; it plays a particularly prominent role within the Social-Cognitive Theory and the works of Albert Bandura [5]. The proponents of observational learning argue that learning by directly experiencing consequences of one's actions is too costly, and not sustainable societally; instead, much of human learning happens vicariously, through observing behaviors of others and consequences of these behaviors [5]. In some cases, individuals model behaviors of others even without witnessing their consequences, particularly when observing figures of authority. For example, in Bandura's classic studies of observational learning, children who observed adults interacting with a doll were likely to replicate the behaviors of the adult (a figure of authority, or a natural model for them) even when they were given no instruction to do so

[4]. Specifically, when adults exhibited gentle behaviors, children tended to play with the doll in a similarly gentle way, whereas children who observed adults being aggressive towards the doll, modeled these aggressive behaviors [4]. Other scholars argued that observational learning is largely responsible for diffusion of attitudes and opinions through a culture [21].

Bandura outlined four processes that are foundational to observational learning: attention, retention/memory, initiation/motor, and motivation. First, to be able to replicate behaviors of others, individuals need to *attune* to or recognize salient defining properties of these behaviors. In addition, once these important properties are recognized, individuals need to be able to *retain* them in their memory, particularly at the time of action. Moreover, individuals need to be able to *enact* the behaviors they wish to replicate. Finally, individuals need to have motivations or incentives to *replicate* modeled behaviors; these motivations are usually reinforced through observing others rewarded for the modeled behaviors, or when models represent authority [5].

In this work, we are interested in examining the application of observational learning in the context of crowdsourcing communities. In the vast majority of the contemporary crowdsourcing communities, workers remain isolated from each other, and are largely unaware of solutions provided by others. In this study we examined whether reviewing solutions generated by others can lead to accuracy gains in one's own future work. The design approaches proposed here specifically targeted the four processes necessary to enable observational learning. They supported attention by explicitly highlighting discrepancies between workers own solutions and solutions provided by others; retention and action by allowing workers to immediately change their own solution to match solutions modeled by others; and motivation by showing how many others selected different solutions, thus using crowd as an authority.

### Facilitating Learning in Crowdsourcing Communities

The notion of feedback has been previously explored in the context of crowdsourcing communities. For example, previous studies showed that exposing crowd workers to feedback on their performance has a positive impact on the level of their engagement and participation [15]. Other researchers specifically examined the impact of feedback on the quality of workers' contributions. Dow et al provided crowd workers engaged in writing product reviews with two different feedback mechanisms: self-generated (in which workers could rate their own performance), and expert-generated [11]; in both conditions workers could revise their answers in light of the provided feedback. Both of these mechanisms were found effective and resulted in improved motivation and performance; in addition, self-assessment resulted in significant learning gains, whereas for expert assessment these gains were marginally significant. In addition, Zhu et al showed that evaluating work by others helped Turkers improve their own performance [23]. Moreover, individuals who provided evaluations within interactive teams demonstrated the most substantial improvement.

The approaches to learning examined in these previous studies, self-assessment, receiving evaluation from experts, and evaluating work of others are theoretically sound and appeared effective in the studies. However, they all have a number of limitations. For example, expert-generated feedback requires involvement of external expert who may not always be available in the context of crowd work. Moreover, critical feedback from experts can be demoralizing. On the other hand, both self-assessment and evaluating work of others create additional tasks for crowd workers, may negatively impact their efficiency, and have a direct impact on the costs of task completion.

## DESIGNING LEARNER-CENTERED CROWDSOURCING

We relied on principles for generating feedback in learning environments to design the different feedback mechanisms discussed in this study. Below we describe the different design approaches and the principles used to guide the design.

### Expert-generated feedback

Expert-generated feedback is, arguably, one of the most common mechanisms for providing learners with personalized feedback on their performance. There exists substantial evidence in regards to its positive impact on learner's performance and on learning gains [23],[11].

Previous research on expert feedback suggested that it is most beneficial when it provides not only accuracy assessment, but also an explanation of the correct solutions, and analysis of gaps between the learner's current state and the optimal performance [3],[8].

Expert-generated feedback has been previously shown as beneficial in facilitating learning within crowdsourcing communities. For example, Dow et al showed that receiving expert feedback helped crowd workers to generate higher quality product reviews [11].

In this study, we used expert-generated feedback in the following way (Figure 2): After submitting their own solution, the participants received a comparison of this solution to the gold standard provided by the expert. The comparison was provided for each meal/ingredient combination (each check-box). The workers could see both the expert-provided correct answer, and the indication of the correctness of their own answer through color-coding (green indicated correct answers, red indicated incorrect answers). The feedback was displayed next to the participant's own solution with a possibility for them to make changes. In addition to this comparison, the expert provided comments explaining correct answers for each ingredient in the selected meal.

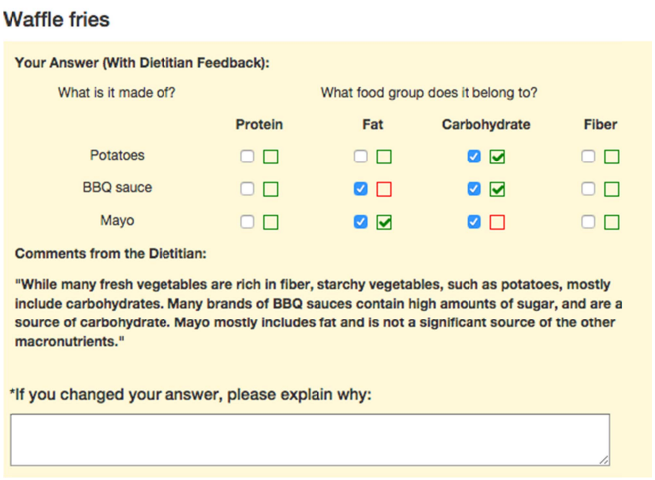


**Figure 2: Expert-generated feedback (green boxes indicate solutions that match the gold standard; red boxes indicate solutions that do not match the gold standard)**

### Explicit peer feedback

Most prior research on peer feedback was conducted in classroom settings where all students both assessed work of their peers and, in turn, received assessments from them. As a result, most of these studies examined aggregated effect of providing and receiving peer feedback. Generally, these studies showed that peer feedback is beneficial; moreover, it has a number of unique benefits as compared to feedback generated by teachers. For example, peer-generated feedback is usually produced at a comparable level of knowledge, and is easier to understand and integrate in one's own work [19], [10]. In addition, while negative feedback from teachers can have a demoralizing effect on students, it appears easier for students to be criticized by their peers. Most importantly, reviewing feedback from peers can expose individuals to different perspectives and help to clarify standards of good performance [20], [8].

In the context of crowdsourcing communities, receiving peer feedback can make workers more cognizant that others evaluate their work and, thus, can affect motivation [3]. Strijbos et al showed that in the context of essay writing assignments, simple feedback from peers had a more positive impact on performance than more elaborate feedback from experts that included error analysis and recommendations for improvement [22]. Gielen et al showed that including justification of peer evaluation can has positive impact on learning [13]. Zhu investigated the effect of assessing the work of others (without receiving any assessments in return) [23].

In this study, we build upon these prior works in the design of the explicit peer-generated feedback conditions. Here, after submitting their own solution, the workers are shown an assessment of the accuracy of that solution by others, coupled with explanations for why certain solutions were deemed incorrect. However, when displaying peer-generated feedback, one of the challenges is addressing possible differences in opinions among peers in their evaluation of correctness of a given solution. We explored two different ways to address these discrepancies in two separate conditions related to explicit peer feedback. In the first such condition (C3 Peer-Feedback/Explicit/MostPopular, Figure 3), the workers were only shown the most popular assessment of their solution (correct/incorrect) provided by the majority of peers.

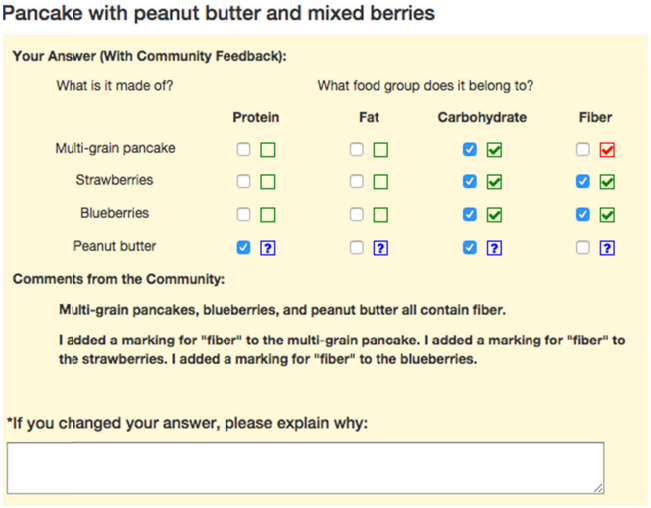


**Figure 3: Explicit peer-generated feedback, most popular assessment.**

In the second peer-feedback condition (C4 Peer-Feedback/Explicit/Distribution, Figure 4), the participants were shown the most popular assessment, together with the frequency of that assessment in comparison to all assessments received (e.g. 4/5).

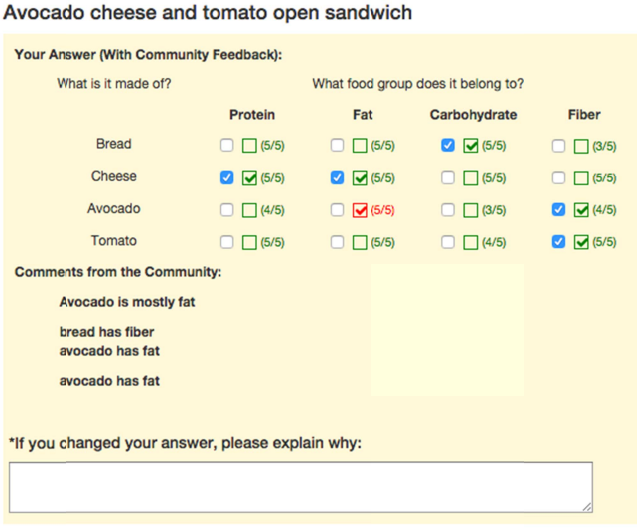


**Figure 4: Explicit peer-generated feedback, distribution of assessments**

### Implicit peer feedback

Previous research suggested multiple benefits of peer-generated feedback. However, generating such feedback

requires introducing additional assessment tasks, and, consequently, may have a negative impact on both time required to complete a crowdsourcing job and on its cost. In this study we explored the possibility of using comparison of one's own solutions to solutions generated by others as an alternative way of generating peer feedback.

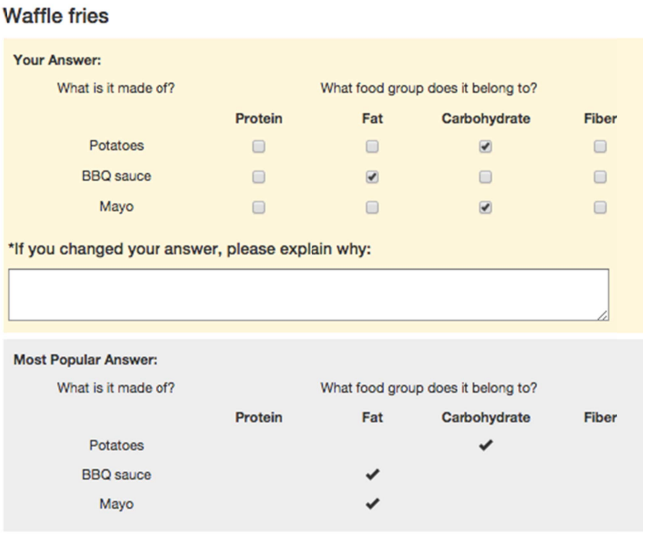


**Figure 5: Implicit peer feedback, showing only the most popular solution**

Analogously to the explicit peer-generated feedback, implicit peer-generated feedback was provided in two different ways. In the first implicit peer-feedback condition (C5, Peer-Feedback/Implicit/MostPopular, illustrated in Figure 5), after submitting their own solution, the participants were presented with the most popular solution provided by other workers. The system did not provide any comparison between the individual's own solution and the most popular solution generated by others; instead it suggested that the individual examines this solution on their own to assess possible differences.

In the second system-feedback solution (C6, Peer-Feedback/Implicit/Distribution, Figure 1 in the Introduction), after submitting their own solution, the participants were presented with a graphical view of the distribution of solutions provided by other workers and their frequencies.

Finally, condition C1, Control, did not include any feedback mechanisms. In this design, the workers were simply asked to provide their solutions one after another. Table 1 summarizes all the experimental conditions.

| Source | Level of Aggregation | |
|---|---|---|
| | Most popular | Distribution |
| Expert | C2 | N/A |
| Peer (Explicit) | C3 | C4 |
| Peer (Implicit) | C5 | C6 |

**Table 1: Experimental conditions (C2-C6, C1-control group, no feedback)**

*Our hypothesis in this study was that all of the experimental conditions (C2-C6) will lead to performance gains across meals and ingredients.*

## METHOD

To explore the research questions above, we designed a controlled study that was conducted with workers on Amazon Mechanical Turk (n=240, between 40 and 45 workers per condition). The basic task was to view digital images of meals with provided ingredients, and map these ingredients to different macronutrients, including protein, fat, carbohydrate, and fiber. We argue that this task fits our definition of knowledge-based tasks because it requires specialized nutritional knowledge.

*Design.* The study used a between-subjects design; the between-group factor corresponded to the type of learning mechanisms as described above (conditions C1-C6).

*Procedures.* Participants were recruited from MTurk, and the tasks were performed as part of their usual paid work. Each task was presented on a separate screen with detailed instructions. The workers were not limited in the time they wished to take for examining the image and completing the task. The workers were asked to complete 20 individual tasks to collect payment for the HIT.

The study was conducted in several phases. In the first *seeding* phase, we recruited 30 workers to provide solutions to the 20 tasks in the dataset without receiving any feedback. These 30 solutions were used as a basis for the implicit peer feedback in conditions C5 and C6. We chose this approach to ensure that all participants in the main experimental phase received consistent feedback. These workers also served as the control group (C1).

In the next, *evaluation* phase, we selected 5 most popular answers for each of the ingredients in the meals, and submitted them for evaluation by Turkers; each meal was evaluated by 5 Turkers (total n=75). Notably, during this evaluation, the Turkers were asked to provide explanation of solutions deemed incorrect. These evaluations were then used as a basis for explicit peer-generated feedback in conditions C3 and C4. This approached meant, however, that some solutions generated during the experimental phase were not evaluated. For these solutions the workers received no feedback.

In the final, *experimental* phase, the workers were assigned to different conditions (C2-C6) to receive different types of feedback. The first 5 tasks in each set (baseline tasks) were used to determine an individual worker's baseline performance; during these 5 tasks workers in all conditions simply completed one task after another with no additional feedback. After completing each of the middle 10 tasks (6-15, training tasks), workers received feedback appropriate for their condition. Finally, the last 5 tasks (16-20, post-training tasks) were used to determine gains in accuracy; during these tasks the participants in all conditions received no additional feedback to maximize consistency between

conditions. To counter-balance for any possible differences in task complexity between different meals, the order of tasks was randomized for all 20 tasks.

*Creating study dataset.* For this study, we used images of meals from Wikimedia Commons available for free download under Creative Commons (CC) license. A professional dietitian analyzed the images, identified components of each of the meals, identified meal ingredients and mapped these ingredients to macronutrients; this mapping was used as gold standard for evaluating Turkers' performance. We selected images that contained meals with 1-5 different ingredients avoiding packaged foods that could present challenge in regards to identifying their content. Importantly, to ensure that we provide workers with opportunities for learning, we selected meals with repeated ingredients. Specifically, we identified 5 "key ingredients"; these included beans, cheese, avocado, nuts, and corn. These ingredients were selected in collaboration with the dietitian on our research team because of common misconceptions regarding their micronutrient content. For example, it is common for individuals to classify beans as protein only and miss their carbohydrate content. Then, for each key ingredient, we found three pictures of meals that included that ingredient (other ingredients varied between these meals). During the study, the participants were exposed to these key ingredients three times: first time during the baseline phase, second time during the training phase during which participants in experimental conditions received feedback on the accuracy of their solution, and finally during the post-training phase with no feedback.

*Measures.* Our primary focus in this study was on the quality of workers' performance, and gains in accuracy achieved during the study. Thus, our main measures were:

***Individual performance***, calculated as the fraction of the individual answers each participant got correct (compared to gold standard established by the professional dietitian). Because each meal ingredient could belong to any combination of food groups (e.g., milk can be classified as protein, fat, and carbohydrate), we captured what fraction of individual macronutrients was correctly identified for each ingredient. For example, if for "beans" the correct answer included "carbohydrates", "protein", and "fiber", and the worker only selected "carbohydrates", their performance score would be 50% (two out of four macronutrients were assessed correctly).

***Accuracy gain***, computed as the difference in accuracy between the last five tasks and the first five tasks. We computed accuracy gains separately for all ingredients in the meals included in these tasks, as well as for the 5 key ingredients because they were repeated across study phases and presented a greater opportunity for learning.

We captured the following subjective measures (each on a 5-point Likert scale):

- ***Self-efficacy gain***: difference between the post- and pre-study response to the question "How confident are you that you can correctly recognize what food groups different foods belong to?"
- ***Subjective difficulty of the task*** ("For this HIT, how would you assess the level of difficulty, on average?", from very difficult to very easy)
- ***Confidence in response*** ("How confident are you that your answers were correct?" from very confident to not at all confident)
- ***Perceived benefits*** of the condition-specific additional information related to learning (e.g., "Seeing answers submitted by other Turkers helped me get better at mapping different foods to their food groups", from strongly agree to strongly disagree. For the participants in the control group C1, this question asked about the benefit of repeating the task multiple times)
- Perceived impact on workers' ***efficiency*** ("All the extra information/feedback I received while working on this HIT made it too time-consuming", from strongly agree to strongly disagree)
- Interest in receiving ***similar feedback in the future*** ("I would like to receive this additional information/feedback in my future HITs" from strongly agree to strongly disagree)
- Perceived ***nutritional learning gains*** in regards to food and nutrition ("After working on this HIT I feel I learned new things about food and nutrition")

Prior to the first task, we recorded Turkers' perceived nutritional literacy ("How knowledgeable are you about food and nutrition?") and an assessment of their self-efficacy.

*Conducting research with MTurk.* The increasing popularity of MTurk as a platform for conducting psychological and sociological research has began to place an undue burden on workers. Many such studies submit HITs where compensation rate does not match time commitment required to complete the tasks. To ensure fair treatment of MTurk workers and following the guidelines for academic requesters, we recorded average time per task for different experimental conditions in a pilot study [24]. We then adopted a compensation rate of close to $10 per hour and used that to estimate the pay-rate for individual HITs.

## ANALYSIS

We used descriptive statistics to explore the dataset and to assess Turkers' baseline accuracy across conditions and for different conditions. We used one-way ANOVA to assess difference in baseline accuracy between different conditions. For the accuracy gains, we used one-sample two-sided t-test comparing mean accuracy gain to zero. To examine difference between participants in regards to subjective assessment measures, including self-efficacy, we used one-way ANOVA, with post-hoc comparisons using Bonferroni correction.

## RESULTS

### Baseline Accuracy Across Workers

We used all solutions submitted for the baseline questions (first 5 questions without feedback across conditions) to calculate the baseline accuracy of Turkers' solutions as the proportion of correct answers to all answers provided. When we looked at per-ingredient accuracy across the different macronutrient/meal/ingredient combinations (how many check-boxes they checked correctly), the participants' baseline accuracy was at 76% (76% of all checkboxes received correct answers across meals and ingredients). The one-way ANOVA test showed that there were no significant differences in accuracy between subjects in different conditions at baseline ($F=0.992$, $p=0.42$).

### Feedback Accuracy

We used solutions collected during the Seeding Phase and the Evaluation Phase as the basis for the explicit and implicit feedback in this study.

#### *Implicit feedback*

We used solutions submitted during the Seeding Phase as the foundation of the implicit feedback in the study. We examined two different ways of providing peer-generated feedback. The first of these conditions, C5 (Peer-Feedback/Implicit/Simple), included only the *most popular solution* generated by Turkers for each ingredient/meal combination. We found that this solution was consistent with gold standard for only 45% of all meal/ingredient combinations (27 out of 60). Across meals, Turkers were consistent in accurately recognizing grain-based foods (such as breads and pastas) as carbohydrates, oils and butters as fat, and cheeses as fat and protein, among others. However, they also exhibited a number of common misconceptions, for example assessing fruits and vegetables as fiber only, rather than as carbohydrate and fiber.

The most popular solutions for the key ingredients, which were repeated across meals, are presented in Table 1.

| Ingredient | Gold Standard | 1st popular | 2nd popular | 3rd popular |
|---|---|---|---|---|
| Beans | CrFbPr (21%) | **Gold standard** | Pr (14%) | Fb (9%) |
| Avocado | FbFt (8%) | Fb (22%) | Pr (16%) | Ft (13%) |
| Nuts | CrFtPrFb (4%) | FtPr (37%) | Pr (17%) | CrFtPr (16%) |
| Corn | CrFb (40%) | **Gold standard** | Fb (22%) | Cr (11%), Pr (11%) |
| Cheese | FtPr (46%) | **Gold standard** | CrFtPr (19%) | Ft (14%) |

**Table 2: Gold standard and the most popular solutions for meals in testing set (Cr=Carbohydrates, Fb=Fiber, Pr=Protein, Ft=Fat). When the most popular solution corresponded to gold standard, we include "Gold Standard" in the table.**

As one can see from this table, these ingredients presented several different scenarios. For two of these ingredients, corn and cheese, the most popular solution generated by Turkers overlapped with gold standard; it was selected by 40% of workers for corn and 46% of workers for cheese. For avocado and nuts, however, the gold standard was not among the three most popular solutions; it was selected only by 8% of workers for avocado, and by 4% for nuts. Both of these ingredients include a complex combination of macronutrients, and most solutions generated by workers missed either one or several of the macronutrients in their solutions. For beans, the opinions of workers were evenly split between the gold standard, which was selected by 21% of workers and another solution, selected by another 21% of workers, who correctly identified Fat and Protein, but missed Carbohydrate.

In the second of the implicit peer-feedback conditions, the workers were provided with a distribution of frequencies for all solutions generated by peers in a graphical form. In this case, we attempted to estimate the accuracy of the aggregated solutions across multiple workers and assess whether this aggregated solution approximated the gold standard. In such an aggregated solution, if 5 workers classified beans as only "Protein", another 5 as only "Fiber" and another five as only "Carbohydrate", their aggregated solution would include all three of these macronutrients, a correct solution, with a popularity count of 5 (100%). However, because for most meal/ingredient combination, each macronutrient received at least one and often more votes, it became necessary to establish threshold of popularity at which a macronutrient would become considered included in the aggregated solution. For example, in the scenario above with "beans" as an ingredient, the question would be whether "fiber" should still be included in the aggregated solution if it received only 3 votes, as opposed to 5. After some experimentation, the final popularity threshold was set to 30% (at least 30% included that macronutrient in their answer).

#### *Explicit feedback*

We used workers' assessments of the top 5 solutions generated during the Seeding Phase as the foundation of the explicit peer feedback.

First, we examined the accuracy of individual assessments submitted during the evaluation phase across meals and ingredients. Across out dataset, the accuracy of an individual assessment was 70% (somewhat lower, but comparable with Turkers' accuracy when simply providing their own answers, which was 76%).

Next we examined the accuracy of these assessments aggregated across 5 Turkers that were used to generate peer-feedback for conditions C3 (Peer-Feedback/Explicit/Simple) and C4 (Peer-Feedback/Explicit/Detailed). Across all meals and ingredients these assessments were accurate 81% of the time. This was significantly higher than their accuracy when they simply provided their own answers ($t=12.29$, $p<0.001$). For the five key ingredients, the accuracy of Turkers' assessments was somewhat lower, and averaged at 78%. The accuracy of the assessments for each of the key

ingredients was the highest for corn (94%) and cheese (92%), lower for beans (85%) and avocado (74%), and the lowest for nuts (63%).

### Accuracy gains

The main question in this study was whether providing accuracy feedback can help crowd workers improve their performance, and what type of feedback leads to the most optimal results. Here we define accuracy gain as the difference in the individual accuracy (number of correct checkboxes) between the first five tasks (baseline) and the last five tasks.

We used a set of one-sample t-tests with Bonferroni correction to assess performance gains for each of the condition, comparing mean accuracy gain to zero. We examined these gains separately across meals and ingredients, and specifically for the key ingredients. The results of this analysis are presented in Table 1.

| Condition | Mean gain across ingredients (t, p-value | Mean gain for key ingredients (t, p-value) |
|---|---|---|
| C1 (control) | -0.59 (-0.52, 0.6) | 0.17 (0.1, 0.92) |
| C2 (expert) | **7.48 (5.69, <0.01)** | **11.1 (6.27, <0.01)** |
| C3 (peer/explicit/simple) | -0.18 (-0.16, 0.87) | 1.19 (0.77, 0.44) |
| C4 (peer/explicit/detailed) | -0.4 (-0.34, 0/73) | -0.49 (-0.28, 0.78) |
| C5 (peer/implicit/simple) | 1.56 (1.25, 0.21) | 1.59 (0.7, 0.48) |
| C6 (peer/implicit/detailed) | **2.98 (2.48, 0.02)** | 3.0 (1.9, 0.06) |

**Table 3: Accuracy gains for different conditions. Conditions with significant gains are marked in bold.**

The performance gains were significant for expert-feedback condition (C2) across meals and ingredients and for key ingredients. The second condition that showed significant performance gains was Peer-Feedback/Implicit/Detailed (C6), however the gains were somewhat smaller for the key ingredients.

### Subjective impressions

In regards to subjective impressions, the study showed significant differences among participants in different conditions for all measures. The participants in conditions C2 (expert-feedback) and C1 (control group) rated the tasks as significantly more difficult than participants in other conditions ($F=10.035$, $p<0.001$). The participants in these conditions (C1 and C2) were also significantly more confident in the accuracy of their solutions than participants in all other conditions ($F=8.6$, $p>0.001$). In regards to their perceived ability to accurately perform similar tasks in the future, the participants in conditions C1 (control), C2 (expert) and C6 (Peer-Feedback/Implicit/Aggregated) were significantly more confident than others ($F=3.09$, $p=0.01$). Interestingly, the participants in the control groups perceived the tasks as significantly more time consuming than all other participants, even though these participants spent less time across the task due to the lack of feedback ($F=5.52$, $p<0.001$). The participants in all feedback conditions reported higher perceived benefit of their condition-specific feedback, as compared to the perceived benefit of repeating the task multiple times for the control group ($F=3.35$, $p=0.006$) and reported higher perceived learning gain as compared to participants in the control group ($F=2.85$, $p=0.02$).

In regards to change in self-efficacy, participants in expert-feedback condition C2 reported the highest gain (mean gain 0.56); the participants in condition C3 (Peer-Feedback/Explicit/Simple) reported loss in self-efficacy (mean gain=-0.05), and the only significant difference was between these two conditions ($F=2.83$, $p=0.17$)

## DISCUSSION

In this research, we set to examine the impact of peer feedback on performance accuracy and learning gains of workers in a crowdsourcing community as compared to no feedback on one hand, and to expert-generated feedback as gold standard. We considered peer feedback in two different forms: explicit, in which individuals received direct evaluations (correct/incorrect) from other workers, and implicit, in which individuals simply compared their own answers to answers provided by others. Many previous studies suggested that peer feedback is a valuable resource and can lead to improvements in motivation and performance. It presents an attractive alternative to the more expensive feedback generated by experts, and to the more time-consuming self-assessment.

The study generated several findings worthy of further explorations.

First, it confirmed that expert-generated correctness feedback with explanations of correct answers is a powerful mechanism for helping crowd workers improve their accuracy on knowledge-based tasks. When such feedback is available, it can help workers to improve their understanding of the tasks, gain necessary knowledge, increase their confidence and self-efficacy, and improve the accuracy of their solutions overtime.

However, it also showed that using solutions generated by other workers as a point of comparison can have a significant positive impact on workers' accuracy and lead to performance gains. This occurred despite the fact that the average accuracy of Turkers' solutions was only at 76% per check-box. We hypothesize that exposing workers to the variety of solutions provided by others and to the relative frequencies of these solutions helped them to consider new possibilities and refine their knowledge. This finding is significant because this form of feedback is readily available, does not depend on availability of experts, does not require introducing additional tasks, and does not lead to increases in workers' workload, time required to compete the crowdsourcing job, and its cost. With simple

modifications to their interfaces, many crowdsourcing communities can incorporate this form of feedback into their repertoire.

The findings also suggested that for such tasks as nutritional assessment of meals, aggregating solutions across individual workers may be a better strategy for arriving at the final solution than identifying the most popular solution. Both of these approaches have been explored in previous research. For example, while PlateMate relies on voting to select between alternative solutions in regards to nutritional assessment of meals [17], Soylent uses both voting and aggregation for such tasks as shortening of text and proof-reading [7].

These findings highlight several important properties of the nutritional assessment task used in our study that we believe contributed to the positive impact of peer feedback. They also suggest a possibility to generalize to a broader class of tasks that might benefit from similar solutions. Specifically, we argue that the task in this study had two essential properties: 1) it relied on a combination of domain-specific knowledge and awareness of existing social norms and conventions; and 2) knowledge in this task was distributed across many individuals who all possess different parts of it. First, mapping ingredients to different macronutrients requires both knowledge of different macronutrients, and also understanding of social conventions in regards to what amount of each macronutrient is relevant for diet management (because strictly speaking, most common foods include all macronutrients, but some in negligible amounts). Second, in mapping ingredients to nutrients, the most common mistake was not to include wrong macronutrients, but to miss some macronutrients for complex ingredients. As a result, putting many partially correct solutions together led to a more complete and accurate aggregated solution. We suggest that other domains that exhibit these properties include coding/programming (e.g. [12]), design critiques (particularly identifying design limitations, e.g. [16]), and copy editing of texts (e.g. [7]). Tasks in each of these domains require a combination of specialized knowledge (e.g. a syntax of a particular programming language) and socially-constructed norms (programing conventions and good practices). We propose that in each of these tasks/domains, exposing workers to solutions generated by others as a form of feedback may enable observational learning and not only contribute to higher quality solutions, but also help workers acquire new knowledge.

On the other hand, the study showed that explicit feedback generated by peers led to decrease in their accuracy. A possible reason for this finding is limitation in the accuracy of the peer-generated assessments. We saw that for three out of five key ingredients, beans, avocado, and nuts, the workers received consistently incorrect feedback, which likely made them question their own knowledge.

The study also suggested many new questions we hope to address with future research. Most importantly, in this study our main focus was on the type of user feedback (explicitly generated by other workers or generated as part of their completion of their own tasks) and on the form in which it was presented (most popular only or with distributions of opinions among peers). To accomplish this, all the feedback in the study was generated in advance as part of the seeding phase and the evaluation phase. In the real world situations, however, this approach is not feasible, and may not be beneficial. Instead, we imagine that the feedback will be generated on the fly and updated with each new submitted solution. This, however, leaves the question of how to scaffold initial solutions, for which no peer feedback is yet available. This also leaves a question of whether incorrect peer feedback early on can have a disproportionate negative impact on the crows-generated solutions. For example, if the first few workers who completed the task provided incorrect answers, would it lead to an information cascade and increase the chance of an incorrect ultimate answer? In addition, all the exploration discussed here focused on nutritional assessment of meals, and specifically on identifying macronutrient composition of different ingredients. Further research is needed to assess whether solutions found beneficial in this study can be generalized to other tasks and domains.

## CONCLUSIONS

In this paper we assessed the impact of peer feedback on crowd workers' performance and learning gains in the context of nutritional assessment tasks. Workers recruited from Amazon Mechanical Turk were asked to match ingredients of meals with corresponding food groups. Some workers were asked to complete 20 tasks in a row with no feedback, others were exposed to different mechanism for facilitating learning, including expert-generated feedback, and two types of peer-generated feedback, explicit and implicit. The study showed that in addition to expert-generated feedback, a comparison of one's own solutions to the distributions of solutions generated by other workers and their comparative frequencies leads to significant improvements in workers' accuracy. We conclude that peer feedback is a powerful mechanism for facilitating learning in crowd computing.

## ACKNOWLEDGEMENTS

This work was supported in part by the NSF grant 1551708, SCH: EAGER: Improving Nutritional Literacy and Decision Making with Learner-Centered Crowdsourcing (PI: Mamykina, L.) and by a Sloan Research Fellowship (Gajos).

## REFERENCES

1. Luis von Ahn. 2009. Human computation. *Proceedings of the ACM International Conference on Image and Video Retrieval*, ACM, 1:1–1:1. http://doi.org/10.1145/1646396.1646398

2. Luis von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM, 1–2. http://doi.org/10.1145/2449396.2449398
3. John Annett. 1969. *Feedback and Human Behaviour: The Effects of Knowledge of Results, Incentives and Reinforcement on Learning and Performance*. Penguin Books Ltd, Harmondsworth.
4. A. Bandura. 2001. Social cognitive theory: an agentic perspective. *Annual Review of Psychology* 52: 1–26. http://doi.org/10.1146/annurev.psych.52.1.1
5. Albert Bandura. 1977. *Social learning theory*. Prentice-Hall, Oxford, England.
6. Albert Bandura. 2001. Social Cognitive Theory of Mass Communication. *Media Psychology* 3, 3: 265–299. http://doi.org/10.1207/S1532785XMEP0303_03
7. Michael S. Bernstein, Greg Little, Robert C. Miller, et al. 2010. Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM, 313–322. http://doi.org/10.1145/1866029.1866078
8. W. Black and D. Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5, 1.
9. Michelene T. H. Chi. 2009. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science* 1, 1: 73–105. http://doi.org/10.1111/j.1756-8765.2008.01005.x
10. Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learning and Instruction* 20, 4: 328–338. http://doi.org/10.1016/j.learninstruc.2009.08.006
11. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1013–1022. Retrieved July 12, 2013 from http://dl.acm.org/citation.cfm?id=2145355
12. Ethan Fast, Daniel Steffee, Lucy Wang, Joel R. Brandt, and Michael S. Bernstein. 2014. Emergent, Crowd-scale Programming Practice in the IDE. *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2491–2500. http://doi.org/10.1145/2556288.2556998
13. Sarah Gielen, Elien Peeters, Filip Dochy, Patrick Onghena, and Katrien Struyven. 2010. Improving the effectiveness of peer feedback for learning. *Learning and Instruction* 20, 4: 304–315. http://doi.org/10.1016/j.learninstruc.2009.08.007
14. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, et al. 2013. The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 1301–1318. http://doi.org/10.1145/2441776.2441923
15. Tak Yeon Lee, Casey Dugan, Werner Geyer, et al. 2013. Experiments on Motivational Feedback for Crowdsourced Workers. *Seventh International AAAI Conference on Weblogs and Social Media*. Retrieved October 25, 2013 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6118
16. Kurt Luther, Jari-Lee Tolentino, Wei Wu, et al. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 473–485. http://doi.org/10.1145/2675133.2675283
17. Ference Marton, Dai Hounsell, and Noel James Entwistle. 1997. *The Experience of Learning: Implications for Teaching and Studying in Higher Education*. Scottish Academic Press.
18. Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, ACM, 1–12. http://doi.org/10.1145/2047196.2047198
19. S.G. Paris and A.H. Paris. 2001. Classroom Applications of Research on Self-Regulated learning. *Educational psychologist* 36, 2.
20. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2: 119–144. http://doi.org/10.1007/BF00117714
21. D. Schaster, D. Gilbert, D. Wegner, and M. Nock. 2014. *Psychology*. Worth Publishers, New York, NY.
22. Jan-Willem Strijbos, Susanne Narciss, and Katrin Dünnebier. 2010. Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction* 20, 4: 291–303. http://doi.org/10.1016/j.learninstruc.2009.08.008
23. Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 1445–1455. http://doi.org/10.1145/2531602.2531718
24. Guidelines for Academic Requesters - WeAreDynamo Wiki. Retrieved September 22, 2014 from http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters