Xinru Tang; Dongyang Zhao; Ying Zhang; Xianghua Ding (2019): AuDi: an Auto-Feedback Display for Crowdsourcing. In: Proceedings of the 17th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing an the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2019_ep05

AuDi: an Auto-Feedback Display for Crowdsourcing

Xinru Tang, Dongyang Zhao, Ying Zhang, Xianghua Ding
Shanghai Key Laboratory of Data Science, Fudan University; School of Computer Science, Fudan University
{15300160005,15307110394,16210240023,dingx}@fudan.edu.cn

Abstract. While feedback, by experts or peers, is found to have positive effects on crowdsourcing work, it is a costly approach as more people or time is involved in order to provide feedback. This paper explores an automatic feedback display called AuDi for crowdsourcing. AuDi shows the worker's accuracy rate, which is automatically calculated with the use of an accuracy algorithm, by changing the background color of the task page. We conducted an experimental study with AuDi in the field, and employed both quantitative and qualitative methods for data collection and analysis. Our study shows that, without introducing new cost, such an auto-feedback display is well received by our participants, gives them assurance and more confidence, and also positively contributes to work performance by pushing them to study more and understand better the task requirements.

1 Introduction

Work performance – particularly in terms of quality output, and work experience are common concerns for crowdsourcing. Many factors could lead to quality issues in crowdsourcing, including unqualified workers (Rzeszotarski and Kittur, 2011; Gadiraju et al., 2015), misunderstanding of requirements (McInnis et al., 2016; Kulkarni et al., 2012; Ipeirotis et al., 2010), and so on. A variety of quality control mechanisms have been explored, such as redundancy and majority voting (Callison-Burch, 2009; Franklin et al., 2011), adding test questions to obtain

accuracy of the workers' answer (Liu et al., 2012), using algorithms to infer the true answer such as Bayesian theory or Expectation Maximization (Liu et al., 2012; Ipeirotis et al., 2010) and etc. Some platforms such as AMT simply reject unqualified work after all the tasks are completed, which, however, causes a series of negative effects on workers' experience (McInnis et al., 2016).

In recent years, feedback as a way to enhance crowd work experience and improve quality output has been investigated (McInnis et al., 2016; Dow et al., 2012). Research has shown positive effects of feedback for crowd work. For instance, a study employing self-assessment and expert reviews as feedback illustrates that, "timely, task-specific feedback helps crowd workers earn, persevere, and produce better results" (Dow et al., 2012). While generally positive, however, most of these are based on personal feedback, which may cause an increase of cost as it relies on more people to spend time on giving feedback.

In this paper, we are investigating an approach which provides automatic feedback to the crowd workers in a timely fashion. More specifically, an accuracy algorithm, based on an accuracy calculation method (Feng et al., 2014), is employed and its accuracy result is shown as the background color of the task page in real time as an ambient feedback display. We refer to this ambient automatic feedback display as AuDi in this paper.

With the study, we found that AuDi was positively perceived and well taken into their crowd work. Both qualitative and quantitative results show that AuDi enables participants to know better of their own performance, feel more in control, and enhance their confidence.

2 Related Work

2.1 Quality Control in Crowdsourcing

Crowdsourcing relies on workers' good performance to produce high-quality output. However, since workers involved are from different countries, with different ages and educational levels, their subjective awareness and background knowledge would inevitably affect their understanding and interpretation of task requirements(Ross et al., 2010; Martin et al., 2014; Gadiraju et al., 2015; Ipeirotis et al., 2010). In consequence, the output quality is barely satisfactory, leading quality evaluation and control to be big issues in crowdsourcing (Kittur et al., 2013).

Many algorithms are proposed to measure the quality of submitted answers. The most common method is redundancy and majority voting, in which the answer given by the majority workers is taken as the correct answer (Callison-Burch, 2009; Franklin et al., 2011; Kulkarni et al., 2012; Little et al., 2010). Further, redundancy can not only be used to determine the correct answer but also help to evaluate the accuracy of each worker (Ipeirotis et al., 2010).

Some research has applied workers' accuracy to the estimation of the results, by integrating workers' answers and their accuracy to infer the correct answer. For

example, one strategy is based on the Expectation Maximization (EM) algorithm, which calculates workers' accuracy by using the confusion matrix (Ipeirotis et al., 2010). This method obtains high-quality results but is at the expense of long inference time.

Besides these underlying quality control algorithms, mechanisms are also explored to change the workflow as a way to enhance work performance for quality output. For instance, Wiseman et al. experimented with inserting an additional check stage, however, their results showed that this would not reduce the error rate, because only a check stage does not make people bother to check their answers (Wiseman et al., 2013). Sandy J. J. Gould et al. studied the effect of a lockout in a data-entry task, and similarly the research shows that the lockout mechanism does work in a laboratory setting, but not in the field where people will do other tasks during the lockout period, making lockouts no longer effective (Gould et al., 2016).

2.2 Work Experience in Crowdsourcing

In recent years, not simply quality output, but the quality of work experience has also become a concern for crowd work. As mentioned, rejecting unqualified work is commonly adopted for quality control, however, work is usually rejected by the requester without giving reasons. This is problematic since payment is the primary motivation of workers (Janine, 2016; Brewer et al., 2016). Past research showed that many workers reported not being paid for adequately completed tasks (McInnis et al., 2016; Irani and Silberman, 2013). Users express their concern about submitting unqualified work, and they are also worried that they may not understand the task which would lead to the failure to get the pay (Mao et al., 2013). This also leads to general feelings of unfairness around rejection, since the requester can get access to all the information about the user, while the users know nothing about their performance and the job criteria (McInnis et al., 2016).

To improve crowd work performance and experience, some particularly focus on providing feedback in real time. For example, Dow et al. (Dow et al., 2012) studied different feedback mechanisms, including peer review, expert review and self-assessment, and found that both self-assessment and feedback from outside will significantly increase the work quality. Concerned with risks in user experience caused by reasons such as unclear evaluation criteria, Brian McInnis et al. (McInnis et al., 2016) suggested automated feedback, which will enable the user to know their performance in time, so it can help build trust between the users and requesters, protect honest users from honest error, and meanwhile punish bad actors.

3 Method

Our study was based on a crowdsourcing platform named ZhongYan, which was set up by our lab for crowd work research projects. For the study, we carefully chose

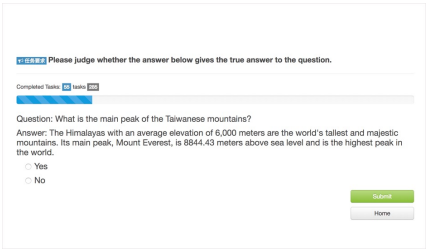

Figure 1. An example of the tasks.

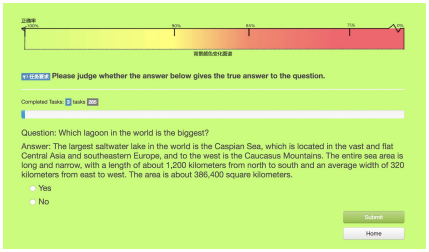

Figure 2. An example for the experimental interface.

and designed crowd task, accuracy algorithm, feedback display format, as well as the experiments, which we will elaborate below.

3.1 Task

A set of text annotation tasks from a real-world project was chosen for our study. For each task, workers are shown one question along with one answer and their job is to determine whether the correct answer appears in the answer text by choosing "yes" or "no" on the task page. We chose it because it was representative of typical crowd task, and was of medium difficulty, which means some may come across a question beyond his or her knowledge, yet he or she can get over this by making extra efforts, for example, Googling. Figure 1 shows an example of the task. In the experiment, every participant was asked to finish about 250 tasks.

3.2 Accuracy Algorithm

For accuracy algorithm, we chose and adapted a quality evaluation algorithm of crowd work proposed by J. Feng et al. (Feng et al., 2014). It is based on Majority Vote (MV), a very popular method to infer the final results in crowdsourcing, and further improves the inference results by considering the different qualities for each worker. This method was chosen because it can achieve a good balance between calculation accuracy and response time compared to other methods (Raykar et al., 2010; Ipeirotis et al., 2010), as it uses an incremental rather than iterative strategy to update the workers' quality. Two models are used in this incremental algorithm. One is the worker model and the other is the question model. The worker model is a quadruple and each element in the quadruple is presented as c_{ij}(i means the answer given by the worker and j means the true answer to the question).

$$\begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix}$$

And the accuracy rate of each worker is calculated as:

$$acc = \frac{c_{00} + c_{11}}{c_{00} + c_{01} + c_{10} + c_{11}}$$

While the question model is presented as a tuple (p_i, 1-p_i) in which p_i is the probability that the true answer to the question is the first choice. And if p_i > 1- p_i, it takes the first choice as the true answer. We build the worker model for each worker to compute the accuracy of the worker and build the question model for each question to infer its result. Each time a worker submits his/her answer we will incrementally update the question model, and the worker model will be updated when we decide the answer to the question in order to acquire the worker's accuracy timely. For our experiments in particular, we used twenty test questions with which we know the correct answers to initialize the worker model of each worker. When it comes to the official questions which lack the correct answers, we compute the question model for each question in order to infer the correct answer combining the submitted answer and the submitter' accuracy computed by his worker model. And we updated each worker's worker model according to the calculated result in order to compute the worker's accuracy timely.

3.3 The Auto-Feedback Display

We decided to show the accuracy information in an ambient form (Mankoff et al., 2003), as it is suitable for persuasion without obtrusion. More so, as representing feedback via color of surrounding area is found to be easier to process and use in goal-striving processes than factual feedback (Ham and Midden, 2010), we decided to use the background color of the task page as a way to show the accuracy information.

As such, we implemented a display schema altering background color of the task page based on accuracy. The color schema is inspired from traffic lights, with red standing for dangerous status, yellow for warning and green for safety. Every time a worker submits his or her answer, the website will change its background color according to the newly calculated accuracy rate while loading the task page.

Through a pilot study on the tasks, we correspond the color schema with an accuracy range from 75% to 100%, so that participants could easily see the change of color while working on the tasks. That is, the accuracy rate of 75% corresponds to the reddest color, and the accuracy rate of 100% corresponds to the greenest. For the sake of convenience, this color schema was shown as a bar on the task page, shown in Figure 2 to help people understand the meaning of the background color.

We chose an accuracy rate of 80% as the acceptance rate for the task. To simulate the real world situation, the participants were told that only those who completed all tasks with an accuracy rate over 80% would earn 50 RMB, otherwise, they would not get paid. Besides, they were also told that they could terminate the experiment anytime they want but only those who finish it in the scheduled time would get paid. After the end of the experiment, however, all got paid as a compensation for their participation in the study.

3.4 Experiment

A total of 50 participants were recruited for the study. These participants were evenly divided into control group and experiment group. Some participants did not show up during our experiments, so at last, there were 22 participants in the control group, 22 in experiment group. We gave participants two days to complete all the tasks, so they could choose any time they like and pause whenever they want, as a way to simulate the real world situation.

Workers in the control groups did not have any feedback - that is, the web background color stayed white (Figure 1), while workers in the experiment group was provided with our automatically calculated accuracy rate as feedback as mentioned above (Figure 2). Participants in the experiment groups were informed of the basic idea of accuracy algorithm in use, and the feedback display. To make it closer to a real-world project, we adopted a redundancy of 5. That is, we divided workers in the experiment groups to subgroups of 5 to calculate accuracy rate within each subgroup. The back-end of the platform recorded each participant's answers and work time for later analysis.

After they finished the project, all workers were assigned an online questionnaire the minute they finished the project to report their self-assessment and personal experience. Participants in the experiment groups were also asked to answer questions about their experience regarding the feedback display while the control groups didn't need to. Almost all the questions of the questionnaire were given in Likert 5-point, except for one question asking participants to write their expected accuracy. Questions covered concentration, confidence, expected accuracy, perseverance and so on, and these data would be for quantitative analysis.

We also conducted interviews with the experiment groups. We recruited 10 interviewees before the experiment started and sought out one more who quit after doing 14 tasks after the experiment. Detail information of these interviewees are listed in Table I ('P' denotes the participants).

ID	Accuracy Rate(%)	ID	Accuracy Rate(%)
P1	86.7	P2	93.3
P3	92.6	P4	89.8
P5	88.8	P6	87.0
P7	88.8	P8	91.9
P9	93.0	P10	90.2
P11	78.6		

Table I. Interviewees' Information.

All the interviews were conducted online through text chat. Each interview lasted about 30 minutes, in which each interviewee reported their experiences about the feedback display and any trouble they came across during the experiment. During our interviews, our questions mainly focused on how they felt

about the feedback display, whether it accorded with their own estimation, how they were possibly influenced by the feedback display, and whether they liked to have the display or not, and why.

3.5 Data Analysis

For quantitative analysis, we collected data from the back-end of the platform and calculated drop out rate, pass rate and accuracy rate. Specifically, we calculated mean, median and standard deviation of these measures for comparison. Meanwhile, we collected results from the questionnaire regarding their estimated accuracy, their confidence level, and so on. For qualitative analysis, we went through the interview data, and identified themes emerged from it. We paid particular attention to those themes that are related to the quantitative results we found from the analysis.

4 Results

Overall, almost all our participants from the experiment groups perceived the feedback provided by AuDi as well reflected their performance, and would all preferred to have it for their crowd work. They further reported that the feedback display had positively influenced their performance and experience, e.g. the red color made them pause for thought and the green color encouraged them to continue. In this way, participants adjusted their work pace accordingly. Below, we will present our results of how people perceived the feedback provided by AuDi, and how AuDi had effects on their performance and experience of doing crowd work.

Group	Completed	Dropped out	Total
Experiment Group	19	3	22
Control Group	21	1	22
Total	40	4	44

Table II. Drop Out Rate.

Group	Passed	Failed	Total
Experiment Group	19	0	19
Control Group	20	1	21
Total	39	1	40

Table III. Pass Rate.

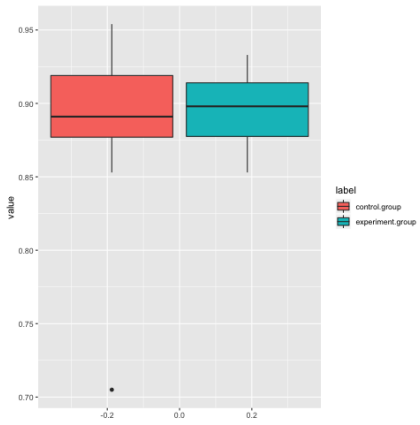

Figure 3. Accuracy Rate.

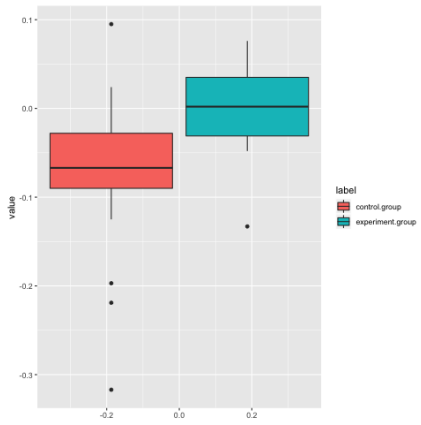

Figure 4. Difference between Self-Estimated and Real Performance.

4.1 Feedback Accuracy and Acceptability

As for the accuracy of the feedback itself, our data shows that participants took the feedback provided by AuDi as largely reliable and quite acceptable. According to the results from the questionnaire, the majority of participants considered the feedback was consistent with their own estimations. Specifically, 26.83% chose level 3, 51.22% level 4, and 9.76% level 5. Our interview data further shows that they considered the feedback was in an acceptable range. For example, P8 commented:

> "Since 100% reliability is unrealistic, I just need feedback stable enough to help me develop a general sense of direction. I mean, the more information provided, the more helpful."

4.2 Effects on Work Performance

Overall, we found the experiment groups performed better and more stably (with less fluctuation of accuracy rate) than the control groups (control group SD_1=0.048, experiment group SD_2=0.029). As shown in Table III, the only one who completed the experiment but did not meet the accuracy rate bar (80%) was from the control group. In addition, Figure 3 shows that the median accuracy rate of the experiment group is higher than the control groups (M_1=0.891, M_2=0.898). Besides, it also shows that the only outlier was from the control group: it was 0.748.

Our interview analysis further suggests that the performance difference between experiment and control groups had something to do with the different levels of understanding of the task requirement, with or without AuDi. Many of our participants reported that they did not understand what they were asked to do, although the description and requirement of the task had been given before the experiment, and few people would bother to read the long description of requirements carefully before the task. The use of AuDi pushed participants to

reread the requirement descriptions when they got relatively negative feedback, which, to an extent, enhanced their understanding and so was their performance.

In our interviews, participants from the experiment group reported how AuDi helped them to learn and grasp the knack to solve the task in practice. P3 described to us:

> "At first, I did not figure out the true requirement of this task. However, the screen changed red abruptly after I gave the wrong answers. Then, I read the question and text again, and finally understood the goal of the task. It had gone well since then."

P8 described how he adjusted his ways of doing task according to the timely feedback:

> "There was a time when the color suddenly turned red. This made me realize that my method might be wrong. Then I gradually adjusted my methods according to the variation trend of background color. After several attempts, I finally got the idea of the task. "

Apparently, the feedback, especially when indicating negative results, did make participants pause to think and study more.

4.3 Effects on Work Experience

The feedback of AuDi had even more impact on work experience. They reported how it provided them a way to evaluate their work on their own and adjust themselves accordingly, making them feel more in control and assured. At the same time, the change of color also easily evoked emotional responses from them, helping them to engage with the task or decide to quit eventually.

There is a big difference in participants' estimation of their own accuracy rate between the experiment group and the control group. That is, the control group's self-estimation was significantly lower: according to the questionnaire data, the average estimated accuracy rate given by the experiment group is 89.93%, while it is 79.56% by the control group. After subtracting estimated accuracy rate from the real accuracy rate, we got Figure 4. From this figure, we noticed that the estimated rate was more consistent with the real performance in the experiment group than the control group. As a matter of fact, there were several outliers in the control groups who unnecessarily considered their performance fairly poor: three between 0.41 and 0.6, and one even at 0.21, while all of these four actually performed far better. Our interviews also illustrated that with AuDi, the experiment group had much higher assurance of their performance than the control group. That is, with real-time feedback, AuDi eliminated their feelings of uncertainty or insecurity to a great extent. P10 shared her experience:

> "There were several questions that I found hard to judge. Without the feedback, I wouldn't be able to determine whether I made the right choice or not."

In addition, our participants reported that AuDi made them engaged with the task more. For instance, P1 described:

> "The feedback system helped me a lot to assess my work. I find it provides valuable information for my reference because it somehow interacts with my own thought and urges me to think about the standard to determine right or wrong."

More so, the auto-feedback display also evoked emotional responses, further urging or encouraging workers to find the right direction for doing the work. In our interviews, some reported that seeing low accuracy rate from the feedback display put pressure on them and evoked negative feelings such as anxiety or frustration, which, then, pushed them to work harder in trying and making the right decision so as to lift up the accuracy rates. One example was from PL:

> "The red color made me feel frustrated a little bit and urged me to make sure the next answer is right to change that situation."

However, interestingly, while they reported negative emotional responses when seeing negative feedback, they at the same time expressed positive feelings towards the feedback display. P8 put it this way:

> "When I saw it was red on my screen, I kind of felt relieved. Simply knowing that there was a mechanism detecting my potential errors made me feel secure and urged me to answer prudently. It's like only when you touch the 'bottom' line, can you learn to climb upwards easily."

On the other hand, if the color was always green, indicating fairly good work performance, they would feel more at ease and confident, as expressed by P3:

> "When the color was green or buff, I know I am good enough to get paid, and my worries and anxieties were gone and I would speed up prudently."

However, for those who couldn't find the right direction after several trials and errors, seeing redness all the time also pushed them to quit the project all together. For example, P11 who quit eventually reported:

> "The full-screen redness made my heart uncomfortable. In the following 5 problems, the screen didn't turn any greener. I felt that I couldn't raise the accuracy rate since the questions were totally beyond me. So I decided to quit."

Other participants also revealed that they would quit if they saw the red color all the time. Specifically, when the screen stayed red for several questions, indicating that the accuracy rate was below 80% , as such they thought that they would never understand the task requirement, let alone getting paid, so they were thinking of

quiting. After several tries, the color turned green again, so they decided to continue. This explains why the drop-out rate is slightly higher in the experiment group than that of control group, as they were sure they couldn't meet the requirement. As Table II shows, 3 participants of experiment group chose to quit, while only 1 in the control group dropped out at last.

People's emotional responses may also have something to do with the particular design of the feedback display. That is, the large area on the screen used to display color inevitably drew people's attention and somehow created a sense of immersion, which made them sensitive to the color change and the color itself. Besides, with the color schema of traffic lights, the related color did the right work to draw people the awareness and response, red for urgency and heightened awareness, green for relief and so on. That is, the color display provided participants with a rough but instant notion of the accuracy rate, leading to corresponding responses in a straightforward manner.

5 Discussions

As shown in our findings, AuDi, by automatically providing feedback in real time, helped engage our participants more and steer them to find the right direction for accomplishing the tasks, which then led to better performance in the end. At the same time, it also helped them feel more assured of the work. Overall, the quantitative and qualitative data analysis shows that the employment of AuDi was very well perceived by our participants, showing that it helped improve their work performance as well as work experiences in crowdsourcing.

Our findings of the use of AuDi indicates several advantages of the auto-feedback approach, compared to other automatic mechanisms, for crowd work performance and experience.

First, as shown in the data, although this auto-feedback approach does not explicitly ask or force people to pause and check, seeing the feedback itself, especially negative feedback, leads participants to actually pause, reread the task requirement, and put more thoughts and efforts to try to get things right. Compared to other intervention mechanisms such as inserting check stage (Wiseman et al., 2013), and introducing lockout (Gould et al., 2016), this auto-feedback mechanism provides more control and more autonomy to the hands of the workers, for them to decide on their own to take actions and do adjustments. As such, it was a more graceful, more humane, and more effective approach to engage workers to do the work right.

More so, the use of AuDi, while helping inform workers to make changes, does not introduce new interruptions, and as such largely protects the flow of work, very important for work performance. Studies show while some intervention mechanisms do improve work quality, the interventions shall be used cautiously as it may interrupt and disturb one's flow of work, and may instead have negative effects on work performance (Gould et al., 2016; Wiseman et al., 2013; Dai et al., 2015; Zhang et al., 2018). In general, workers might have difficulty in resuming to

perform tasks after experiencing an interruption and have to take time to regain focus (Iqbal and Horvitz, 2007) or suffer more stress and frustration in order to re-engage in less time (Mark et al., 2008). That is, with inappropriate interventions, interruption cost incurred in switching attention between tasks and as such would negatively affect workers' performance. When AuDi is considered, by simply providing feedback in the ambient form, without inserting breaks or lockout, it greatly minimizes the disturbing effects, and their flow of work protected.

Finally, the use of auto-feedback also appears to potentially relieve the commonly reported tensions between workers and requesters on crowdsourcing. Crowd workers are usually regarded as inexhaustible and anonymous labors, and were managed as such. The criteria of tasks are defined by the requesters and they have the final say in whether to accept the work and pay for them or not (Irani and Silberman, 2013). In consequence, workers are at risk of work rejection and have no reasonable resources to avoid this wage theft (McInnis et al., 2016; Irani and Silberman, 2013). Andrew Mao et al. investigated the reasons why workers drop out a crowd work and found out the most important reason is workers worried about their submitted answers being rejected (Mao et al., 2013). As such, rejecting their work without any reason or feedback was a commonly complained issue in crowdsourcing, and caused a lot of tensions between workers and requesters.

As shown in our study, the use of AuDi, by feeding the performance information back to the workers, not the requesters, quite successfully addressed workers' concern about quality. As reported by our participants, AuDi made our participants more aware of what they were doing in real time, and helped them make decisions on their own whether to go ahead confidently, pause to find ways to fix things, or to even quit completely. That is, what matters is not whether their work is rejected or not, but the reason of why the work is rejected, and AuDi is certainly helpful in that respect.

6 Limitations and Future Work

Though our study shows very positive results about using AuDi, there are also a number of limitations of the system and the study. First, the accuracy calculation method used makes it only work for those tasks with multiple choice questions, and not other tasks. So for crowd work that does not meet this requirement, AuDi can't apply without necessary adaption.

In addition, the particular algorithm might also lead to cold start effect. That is, AuDi's feedback is based on comparing submitted answers to the estimated right answers, so it relies on the already submitted answers to do the estimation. Owing to that, the first few workers will not get feedback as effective as the later ones do, as there are no other answers yet. Whether increasing transparency (e.g. displaying how many submitted answers on which the feedback is based) might be a good solution to this issues still needs further investigations.

Besides, there is also accumulated effect with the approach of AuDi. That is, the accuracy rate and the corresponding background color will change more dramatically at the beginning. As the number of questions answered grow, the accuracy rate will not be so greatly affected by one single answer anymore, so the color change will become less obvious. This effect was noticed and was also reported by our participants in the study, as they could see more background change at the beginning but not so much towards the end. To address this issue, we might divide all questions into multiple subgroups and to initialize the algorithm every time with each subgroup. But at the same time, people would rely more on the feedback at the beginning as learning is more actively taken by workers at the beginning. So it takes further investigations to find out whether the accumulated effect on the display shall be addressed and how.

7 Conclusions

In this paper, we presented an auto-feedback display called AuDi, as well as an experimental study to investigate how AuDi might work for crowdsourcing. Our study shows that people perceived the automatically calculated accuracy feedback as generally acceptable, and the feedback display itself was helpful for them to engage with the tasks and perform the work better. More specifically, it helped raise people's awareness and leading people to pause for thought and do the work more carefully when seeing red color, and encouraging them to proceed with more confidence when seeing green color. Without introducing new cost, AuDi shares the similar positive effects as personal feedback.

Hata et al.'s study shows that a worker's long-term performance is quite stable, as they usually adopt a particular strategy for completing tasks and will continue to use that strategy without change (Hata et al., 2017). However, as shown in our study, this is only true when there is no feedback for their work. When feedback is provided, as the use of AuDi in our case, changes of strategies for better performance could happen over the process.

*

References

Brewer, R., M. R. Morris, and A. M. Piper (2016): '"Why would anybody do this?": Understanding Older Adults' Motivations and Challenges in Crowd Work'. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. San Jose, California, USA, pp. 2246–2257.

Callison-Burch, C. (2009): 'Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk'. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language*. Singapore, pp. 286–295.

Dai, P., J. M. Rzeszotarski, P. Paritosh, and E. H. Chi (2015): 'And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions'. In: *Proceedings of*

the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. Vancouver, BC, Canada, pp. 628–638.

Dow, S., A. Kulkarni, S. Klemmer, and B. Hartmann (2012): 'Shepherding the crowd yields better work'. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. Seattle, Washington, USA, pp. 1013–1022.

Feng, J., G. Li, H. Wang, and J. Feng (2014): 'Incremental Quality Inference in Crowdsourcing'. In: *International Conference on Database Systems for Advanced Applications*. pp. 453–467.

Franklin, M. J., D. Kossmann, T. Kraska, S. Ramesh, and S. Ramesh (2011): 'CrowdDB: answering queries with crowdsourcing'. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. Athens, Greece, pp. 61–72.

Gadiraju, U., R. Kawase, S. Dietze, and G. Demartini (2015): 'Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys'. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea, pp. 1631–1640.

Gould, S. J., A. L. Cox, D. P. Brumby, and A. Wickersham (2016): 'Now Check Your Input: Brief Task Lockouts Encourage Checking, Longer Lockouts Encourage Task Switching'. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. San Jose, California, USA, pp. 3311–3323.

Ham, J. and C. Midden (2010): 'Ambient persuasive technology needs little cognitive effort: the differential effects of cognitive load on lighting feedback versus factual feedback'. In: *Proceedings of the 5th international conference on Persuasive Technology*. Copenhagen, Denmark, pp. 132–142.

Hata, K., R. Krishna, L. Fei-Fei, and M. S. Bernstein (2017): 'A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality'. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, Oregon, USA, pp. 889–901.

Ipeirotis, P. G., F. Provost, and J. Wang (2010): 'Quality management on Amazon Mechanical Turk'. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. Washington DC, pp. 64–67.

Iqbal, S. T. and E. Horvitz (2007): 'Disruption and recovery of computing tasks: field study, analysis, and directions'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. San Jose, California, USA, pp. 677–686.

Irani, L. C. and M. S. Silberman (2013): 'Turkopticon: interrupting worker invisibility in amazon mechanical turk'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France, pp. 611–620.

Janine, B. (2016): 'Income Security in the On-Demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers'. *Comparative Labor Law & Policy Journal*, vol. 37, no. 3.

Kittur, A., J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton (2013): 'The future of crowd work'. In: *Proceedings of the 2013 conference on Computer supported cooperative work*. San Antonio, Texas, USA, pp. 1301–1318.

Kulkarni, A., M. Can, and B. Hartmann (2012): 'Collaboratively crowdsourcing workflows with turkomatic'. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. Seattle, Washington, USA, pp. 1003–1012.

Little, G., L. B. Chilton, M. Goldman, and R. C. Miller (2010): 'TurKit: human computation algorithms on mechanical turk'. In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. New York, New York, USA, pp. 57–66.

Liu, X., M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang (2012): 'CDAS: a crowdsourcing data analytics system'. *Proceedings of the VLDB Endowment*, vol. 5, pp. 1040–1051.

Mankoff, J., A. K. Dey, G. Hsieh, J. Kientz, S. Lederer, and M. Ames (2003): 'Heuristic Evaluation of Ambient Displays'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA, pp. 169–176.

Mao, A., E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith (2013): 'Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing'. In: *Proceedings of First AAAI Conference on Human Computation and Crowdsourcing*. Palm Springs, California, USA.

Mark, G., D. Gudith, and U. Klocke (2008): 'The cost of interrupted work: more speed and stress'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Florence, Italy, pp. 107–110.

Martin, D., B. V. Hanrahan, J. O'Neill, and N. Gupta (2014): 'Being a turker'. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. Baltimore, Maryland, USA, pp. 224–235.

McInnis, B., D. Cosley, C. Nam, and G. Leshed (2016): 'Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk'. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. San Jose, California, USA, pp. 2271–2282.

Raykar, V. C., S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy (2010): 'Learning from crowds'. *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322.

Ross, J., L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson (2010): 'Who are the crowdworkers?: shifting demographics in mechanical turk'. In: *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. Atlanta, Georgia, USA, pp. 2863–2872.

Rzeszotarski, J. M. and A. Kittur (2011): 'Instrumenting the crowd: using implicit behavioral measures to predict task performance'. In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. Santa Barbara, California, USA, pp. 13–22.

Wiseman, S., A. L. Cox, D. P. Brumby, S. J. Gould, and S. O'Carroll (2013): 'Using Checksums to Detect Number Entry Error'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France, pp. 2403–2406.

Zhang, Y., X. Ding, and N. Gu (2018): 'Understanding Fatigue and its Impact in Crowdsourcing'. In: *Proceedings of 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design*. Nanjing, China, pp. 57–62.