FEEDBACK AND PERFORMANCE IN CROWD WORK: A REAL EFFORT EXPERIMENT

Tim Straub
Karlsruhe Institute of Technology, Karlsruhe, Germany, tim.straub@kit.edu

Henner Gimpel
University of Augsburg, Augsburg, Germany, Augsburg, Germany, henner.gimpel@fim-rc.de

Florian Teschner
Karlsruhe Institute of Technology, Karlsruhe, BaWue, Germany, teschner@kit.edu

Christof Weinhardt
Karlsruhe Institute of Technology, Karlsruhe, Germany, christof.weinhardt@kit.edu

Tim Straub, Henner Gimpel, Florian Teschner, and Christof Weinhardt, 2014, "FEEDBACK AND PERFORMANCE IN CROWD WORK: A REAL EFFORT EXPERIMENT", Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, ISBN 978-0-9915567-0-0
http://aisel.aisnet.org/ecis2014/proceedings/track07/6

FEEDBACK AND PERFORMANCE IN CROWD WORK: A REAL EFFORT EXPERIMENT

Research in Progress

Straub, Tim, Karlsruhe Institute of Technology (KIT), Germany, tim.straub@kit.edu

Gimpel, Henner, University of Augsburg, Germany, henner.gimpel@fim-rc.de

Teschner, Florian, Karlsruhe Institute of Technology (KIT), Germany, florian.teschner@kit.edu

Weinhardt, Christof, Karlsruhe Institute of Technology (KIT), Germany, christof.weinhardt@kit.edu

Abstract

Online labor markets gain momentum: Frequently, requesters post micro-tasks and workers choose which tasks to complete for a payment. In virtual, short-lived, and commonly one-shot labor relations, one challenge is to properly incentivize worker effort and quality of work. We present a real effort experiment on a crowd work platform studying the effect of feedback on worker performance. Rank order tournaments might or might not disclose a worker's current competitive position. One might expect that feedback on the competitive position spurs competition and, in effect, effort and performance. On the contrary, we find evidence that in rank order tournaments, performance feedback tends to have a negative impact on workers' performance. This effect is mediated by task completion. Furthermore when playing against strong competitors, feedback makes workers more likely to quit the task altogether and, thus, show lower performance. When the competitors are weak, workers tend to complete the task but with reduced effort. Thus, providing performance feedback might not be advisable in crowd labor markets.

Keywords: Crowdsourcing, Online Labor, Performance Feedback, Rank Order Tournament, Real Effort Experiment.

1 Introduction

Paid crowd work offers remarkable opportunities for distributed work and improving productivity. Moreover it enables a distributed workforce to complete complex tasks on demand and at scale. Crowd work today spans a wide range of skill and pay levels, with commercial vendors (e.g., Amazon Mechanical Turk (MTurk for short), oDesk, Clickworker) providing access to a range of workers and focused support for various tasks. These tasks range from simple repetitive e-mail tagging to creative and complex tasks such as building logos or websites (cf. Kittur et al., 2012, 2013).

In virtual, short-lived and commonly one-shot labor relations, one challenge is to properly incentivize worker effort and quality of work. Quality control is frequently done through repetition of work and managing an individual worker pool (Ipeirotis et al., 2010; Kokkodis and Ipeirotis, 2013; Wang et al., 2013). Incentives typically comprise the payment of a flat fee when the work is acceptable and an additional bonus when the work is very high quality. This raises the question of how incentive schemes can be designed to motivate workers to provide their best effort and deliver high task

performance. Rank order systems are pretty common in work places (Microsoft, GE, Yahoo! etc.) and in competitive environments (Poker, soccer leagues etc.) suggesting that they indeed induce the spirit to deliver higher performance. Hence, it seems reasonable to transfer this common incentive design to crowd work. The research question we explore in this paper is how performance feedback in rank order tournaments (ROTs) among workers affect their effort and task performance.

In this paper we report a real effort experiment on MTurk studying the effect of performance feedback on worker effort in ROTs. In ROTs, on average there is no difference whether feedback is shown or not. In a nutshell, the root for this unintuitive result is participant heterogeneity. While low performers stop working all together, high performers knowing that they will be rewarded work less.

2 Background and Research Model

2.1 Crowd Work

Crowdsourcing and online labor markets have emerged as new labor pools of freelancers that allow organizations to flexibly scale their workforce and hire experts. Today, MTurk dominates the market for crowdsourcing micro-tasks that are trivial to humans but challenging to computers (Ipeirotis, 2010). Examples include tagging images, transcribing audio recordings, verifying addresses and phone numbers. More sophisticated work like software development, design, and innovation contests is performed in online labor marketplaces like oDesk, Clickworker, and Innocentive. Recently, online labor markets (especially MTurk) gained widespread interest as a platform to run low cost experiments with subjects from a demographically diverse pool. Previous work has examined its validity, costs (e.g. Chilton et al., 2010), and participant demographics (Paolacci et al., 2010; Berinsky et al., 2012). See e.g. Mason and Suri (2012), Horton et al. (2011), Kaufmann et al. (2011), Pilz and Gewald (2013), and Teschner and Gimpel (2013a, b) for recent examples.

Two of the main issues with crowd work are (1) how to secure quality and (2) incentivize workers to give their best (e.g. Wang and Ipeirotis, 2013; Shaw et al., 2011). Manual verification of work quality is typically not feasible. Thus, some malicious workers take advantage of the system by quickly submitting low quality work (Ipeirotis et al., 2010). To compensate for low quality work one of the main strategies used is repetition of work. This is, however, costly. Kokkodis and Ipeirotis (2013) show evidence that it is possible to predict a workers performance by categorizing tasks and using feedback. An even stricter approach applied by some requesters is to build one's own trusted workforce with workers who delivered high-quality work in previous requests. Shaw et al. (2011) show that using a combination of social and monetary incentives leads to better quality. Furthermore, pure monetary incentives, such as tournaments, lead to higher performance, but were not significantly different from control conditions. This raises the demand of further investigation. Paolacci et al. (2010) report that compared to laboratories, crowdsourcing needs rather small monetary incentives to get comparable results. This indicates that a good selection of incentives fosters quality in crowdsourcing settings. Contrary, Mason and Watts (2009) find that more money leads to more effort, but quality stays the same. Furthermore, they find that a quota pay scheme, which only pays for a set of completed tasks, leads to a greater output than a piece rate (pay every task) even though the quota payment was smaller. To sum up, it is an open debate which incentive and information structures are best suited to stimulate worker performance and quality output.

2.2 Rank Order Tournaments

A rank order tournament (ROT) is a setting in which two or more people are ranked according to their performance. Only the top performers win the tournament. Such settings are usually used in sports but as well in some crowdsourcing platforms such as 99designs and by some requesters on MTurk.

Bull et al. (1987) compared effort levels under a piece rate payment scheme and a ROT payment scheme. Their findings show that the average effort levels form a Nash equilibrium, but have a higher variance in ROTs. Similarly, Van Dijk et al. (2001) report that effort levels and variance in tournaments are higher compared to a piece rate. Furthermore, workers with a low ability work harder. These results suggest two hypotheses: First, in a tournament some subjects lose interest if they are far behind and have no chance of winning anymore. Others who are far in front might relax. And some who are close to each other might actually be competing. Second, the variance in ROTs might be induced by risk-aversion. Eriksson et al. (2009a) present experimental evidence that when subjects can choose between ROTs and piece rates, variance decreases and effort levels increase in tournaments. They further find that risk-averse subjects tend to choose a piece rate scheme. This suggests that some people are more motivated by tournaments than others.

Eriksson et al. (2009b) experimentally study the influence of feedback on subjects' effort with piece rate payments and ROTs. Each of these is played with three different feedback rules on relative performance. No feedback, feedback given half way through the experiment, and a continuously updated feedback. They find that on average feedback does not change effort. Furthermore, subjects who are behind make more mistakes under continuous feedback. Interestingly subjects who are behind almost never drop out of the tournament. Eriksson et al. (2009b) argue that the reason could be a social norm to never give up. This effect might, however, be stronger in a laboratory setting than an anonymous crowd labor market. Evidence in this direction is presented by Fershtman and Gneezy (2011): While quitting is often socially stigmatized and subjects often try to avoid it, they find that higher rewards lead subjects to exert more effort, but a higher rate of quitting is observed as well. Finally, Pull et al. (2013) show that in dyadic tournaments where the ability of subjects is heterogeneous, effort levels decrease, because both know that one will win anyway. While when the subjects' abilities are homogeneous, the effort levels will be much higher. In consequence we imply that a live or continuous feedback will inform the subjects about their heterogeneity or homogeneity and will lead to the same effect. Furthermore, if participants get feedback and performed better than expected, they decrease their effort but expect to be better in the future (Kuhnen and Tymula, 2012). On the other hand, people who performed worse than their expectations will increase their effort but reduce their expectations. This implies that showing feedback suggesting that they will lose might improve their performance while feedback that they are winning might lower performance.

2.3 Motivation

A further point which is affected is the intrinsic and extrinsic motivation of participants. Intrinsic motivation refers to doing something because one wants to do it out of pure interest or fun, while extrinsic motivation refers to the motivation of doing something out of external reasons like getting a reward (Ryan and Deci, 2000; Eccles and Wigfield, 2002). In this study we clearly affect extrinsic motivation by using monetary incentives for performing well. We cannot completely exclude that some participants are intrinsically motivated. In settings like MTurk the main motivation for workers is to earn money. Therefore workers are mainly externally motivated. By using a real effort task, which from design has no overall epic meaning, is boring, and does not improve major skills of a person, we tried to exclude intrinsic motivations like altruism, entertainment, and personal development as much as possible. Furthermore Deci et al. (1981) argue that competition decreases intrinsic motivation. More accurately competitive situations where participants feel a pressure to win undermine intrinsic motivation (Reeve and Deci, 1996).

2.4 Research Model

Figure 1 summarizes our research model based on the related work reviewed above: Following the sequential distinction of service quality in structure, process, and outcome (Donabedian, 1980, 2003), a worker's performance is considered as outcome and is hypothesized to by influenced by the work

process and structures. Structural antecedents are classified as individual, crowd, or system level. We believe this structure will prove useful for more extensive conceptualization on the interrelation of crowd labor incentives and quality. Evaluating this belief is future work; here the generic structure is used as frame for a specific causal moderated mediation model.

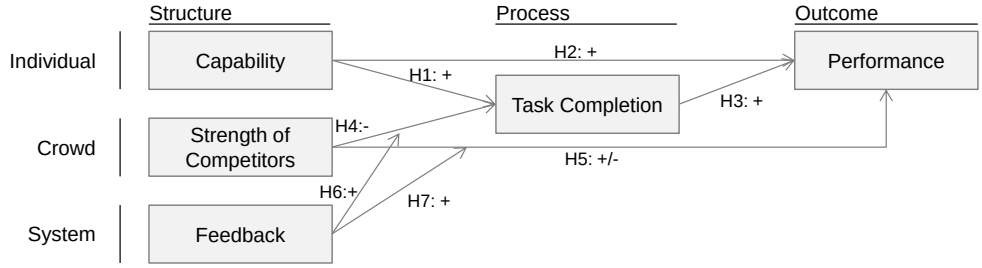

Figure 1. Hypothesized moderated mediation model on the determinants of worker performance in rank order tournaments

Performance is directly affected by the worker's capability, i.e. his ability to perform the specific task. Capability has a positive effect on performance. Performance might be affected by the competitors' strength. Only when feedback on the performance and current standing in a ROT is provided, since it can only be seen by a worker in these cases (cf. Eriksson et al., 2009b). Therefore the effect is moderated by feedback. Given evidence from experiments on ROTs, the direction of the moderated effect of the competitors' strength on performance is, however, not ex-ante clear. Task completion indicates if a worker finished the task or not. In other words, quitting the task is measured. Following prior evidence and common sense, we assume a strong positive effect on performance. We hypothesize that task completion is mediating the effects of capability and competitors' strength on performance. When a worker has the necessary ability to do the task, he is more likely to finish it. Strength of competitors should have a negative effect on task completion: Similar to the findings of Fershtman and Gneezy (2011), knowing that one is falling behind leads to higher rates of quitting. Therefore the stronger the competitor is, the more likely one will quit the task. Feedback is hypothesized to moderate the effect of strength of competitors on both task completion and performance. Only when feedback is given the competitors' strength can be seen and the effects can appear. When competitors are strong, i.e. when a worker faces one or multiple strong competitors, the negative effect of competitors' strength on task completion can be expected to show more strongly than when a worker's competitors are weak. For strong competitors, we hypothesize the effect on performance to be positive while we expect it to be negative for weak competitors. In other words: When a worker sees that he is falling behind but does not quit the task, the feedback is expected to increase performance. When he is ahead, he might relax.

Tournaments might be a fruitful incentive for crowdsourcing settings, but likewise may not work, because crowd workers are used to work under piece rates conditions. Given the partially inconclusive theoretical and empirical evidence related to the hypothesized model, we test it experimentally to evaluate the existence and direction of hypothesized effect in the context of crowd labor markets.

3 Experiment Design and Procedures

The experiment was conducted via MTurk. We followed standard procedures for experimental research. The experiments were conducted with a custom-made web application. From a technical perspective we followed the guidelines of Mao et al. (2012) and Mason and Suri (2012). To measure worker performance, we implemented a real effort task similar to the slider task by Gill and Prowse (2012): workers see a slider ranging from 0 to 100 and have to set it to 50. This is repeated over and over until either the time for the task elapses or workers quit. The rather simple, needless work is by purpose and typical for real effort experiments. The intention is typically to measure workers reaction

to incentives, feedback, and competition, with a simple task that is easy to understand and depends as little as possible on pre-existing knowledge, learning effects, randomness, or guessing (Gill and Prowse, 2012). Furthermore it partially excludes intrinsic motivational factors as already discussed above. The number of sliders a worker correctly sets to 50 prior to the end of the task is the measure of performance. The slider task was originally developed in z-Tree (Fischbacher, 2007). We implemented a similar version in JavaScript, since the experiment should be accessible online through MTurk. In addition, we added an *out* button, to quit the task whenever the workers wanted to. This was used as measure for the binary variable task completion (1 = completed, 0 otherwise). The explicit option to quit the task was intended to reduce experimenter demand effects and the relevance of a potential social norm to never give up.

Subjects were recruited from the general pool of MTurk workers with restrictions that workers reside in the US, completed at least 1,000 tasks on MTurk, and had a task approval rate of at least 95%. Upon agreeing to participate in the experiment, workers received instructions and had to complete a brief quiz testing their understanding. They then worked on the slider task for 1.5 minutes with piece rate payment of USD 0.01 per finished slider. The number of finished sliders under this piece rate is taken as measure for a worker's capability in the task. Next, workers participated in a 3 minute dyadic ROT. In this second round the number of sliders was used as measure for performance. The worker with most finished sliders won the tournament and received a payment of USD 1.00, the other went away empty-handed. In case of a tie, the winner was determined randomly. To increase experimental control and comparability, ROTs were not live but workers played against historic data from a pre-test. This was explained to workers in the instructions. For the ROT, each worker was randomized to either of three treatments: no feedback on the performance of the competitor (NF), feedback on the performance of the competitor in a ROT with a strong competitor (FS), and feedback on the performance of the competitor in a ROT with a weak competitor (FW). The weak competitor always finished 27 sliders while the strong competitor finished 66 sliders. Feedback was measured binary whether it was shown (FW and FS) or not (NF). Finally, workers answered a questionnaire and received their payment for the experiment, i.e. a show-up fee of USD 0.30, a piece rate for completed sliders during the measurement of their capability plus potentially USD 1.00 for winning the ROT.

Figure 2 illustrates the task and the feedback: The left hand side shows an example with feedback (either FS or FW). At any time during the ROT the worker sees his own performance so far (here 7 completed sliders), his competitor's performance so far (here 14 completed sliders), and the next slider to set to *50*. In addition, the screen had a timer and a quit button. The right hand side of Figure 2 exemplifies the NF treatment; it is identical except that feedback on the competitor's performance is missing – it is only disclosed after the ROT when the result is shown.

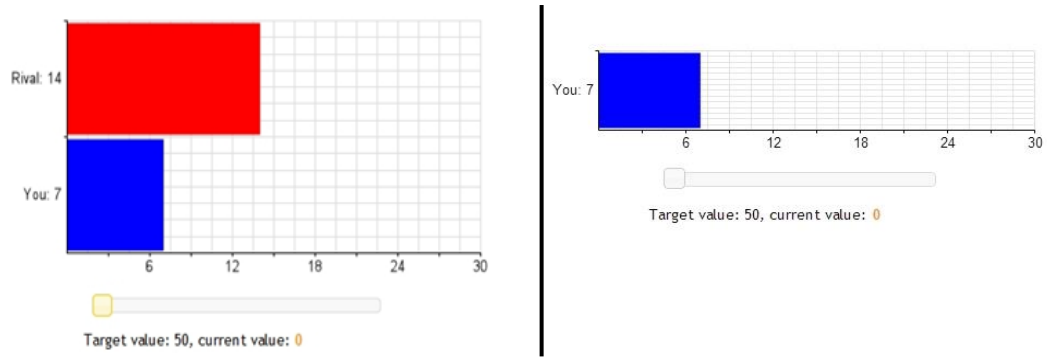

Figure 2. Experiment interface: Live feedback (left image), no feedback (right image)

4 Experiment Results

Overall, 257 workers participated. Table 1 shows summary statistics by treatment. As expected, given random assignment to treatments, there is no significant treatment difference in age (ANOVA, p-value = 0.752), gender (χ^2 test, p-value = 0.97), or average duration of the experiment (ANOVA, p-value = 0.388). Differences between average payment were observed (ANOVA, p-value = 0.001). This is given by experiment design: In FW, the competitor was weak, in FS he was strong. Thus, the rate of workers winning differs. In NF, the rate of weak and strong competitors was equal to the union of FS and FW. In piece rate payment, there is no significant treatment difference in the number of finished sliders per minute, i.e. in workers capability (ANOVA, p-value = 0.632). We conclude that subjects are relatively homogenous across treatments and any performance difference in the ROT is a causal effect of the experimental control over the feedback provided and the competitors' strength.

	Treatment			
	NF	FW	FS	All
Number of participants	97	80	80	257
Age in years (mean, 95% CI)	31.08 [29.53, 32.63]	31.02 [29.61, 32.44]	30.05 [28.61, 31.49]	30.74 [29.89, 31.60]
Share female	40.21%	40.00%	41.25%	40.47%
Duration in minutes (mean, 95% CI)	8.64 [7.93, 9.35]	8.48 [8.00, 8.97]	7.87 [7.46, 8.28]	8.35 [8.02, 8.69]
Payment is USD (mean, 95% CI)	0.95 [0.87, 1.02]	1.34 [1.28, 1.40]	0.49 [0.46, 0.51]	0.93 [0.88, 0.97]
Sliders per minute in piece rate (mean, 95% CI)	11.41 [10.50, 12.32]	12.03 [10.97, 13.10]	11.53 [10.62, 12.45]	11.64 [11.09, 12.19]

Table 1. Descriptive statistics by treatment

We first look at the performance of workers under piece rate payment and in the ROT. We use the number of finished sliders per minute as performance measure. Aggregated over all treatments, performance increased from 11.64 sliders per minute under piece rate payment, to 13.86 sliders per minute in the tournament. The effect is significantly greater than zero (mean difference 2.22; t-value = -5.511; p-value = 0.001). This is in line with the findings of van Dijk et al. (2001). The effect size is medium (Cohen's d = 0.487). It will depend on the application scenario whether such a medium performance increase justifies the extra burden of setting up and communicating a ROT among workers. This result should, however, be interpreted cautiously, as it might be confounded by order effects. Experience and fatigue could corrupt the measurement. Controlling this by balancing the order of incentive schemes was not focus of the experiment.

The moderated mediation model outlined in the previous section is evaluated with a set of 4 regressions, following the general steps from Hayes (2009) contemporary interpretation of Baron and Kenny's (1986) mediation and moderation analysis. We first establish the correlation of the causal variables on the mediator (regression models 1 and 2) and then estimate the effect of causal variables and the mediator on the outcomes variable (regression models 3 and 4). Task Completion is binary (1 = completed, 0 otherwise), so is strength of competitor (weak or strong). In our setting, the statistical consideration of moderation differs from the conventional approach: Conventionally, feedback moderating the effect of strength of competitors would be modeled by two direct effects (one from feedback, one from strength of competitors) and the interaction of these effects. In our model and experiment, strength of competitors is, however, not meaningfully defined in the absence of feedback. Without feedback, strength of competitors cannot affect either task completion or performance. Thus, moderation here results in three combinations: No feedback (irrespective of strength of competitors), feedback and a weak competitor, and feedback and strong competitor. Table 2 provides the results.

	(1)	(2)	(3)	(4)
Dependent variable	Task Completion	Task Completion	Performance	Performance
Estimation method	Logit	Logit	OLS	OLS
Task Completion			21.399 ***	21.399 ***
Capability	0.109 **	0.109 **	1.339 ***	1.339 ***
No Feedback		0.260		-0.771
Weak Competitor x Feedback	-0.260		-2.348 *	-3.119 *
Strong Competitor x Feedback	-1.537 *	-1.277 *	0.771	
Intercept	1.546 *	1.286 +	-1.044	-0.273
N	257	257	257	257
R^2 (Cragg and Uhler's R^2 for logit)	0.164	0.164	0.689	0.689

*Table 2. Regression results (Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1).*

As expected, capability substantially and significantly affects task completion (regression model 1; support for H1). No feedback is a dummy equal to 1 for NF treatment and 0 for FW and FS and is the negation of feedback. The interaction of strength of competitor and feedback assesses the moderation. When competition is weak, and feedback is shown it has no significant effect on task completion compared to no feedback. On the contrary, when facing a strong competitor, it has a significant negative effect on task completion. With a strong competitor, feedback makes workers quit work. The difference between a weak and a strong competitor is significant (regression model 2, significant effect of a strong competitor interacted with feedback). Thus, feedback moderates the effect of strength of competitors on task completion (support for H6) and when it has an effect on task completion, it is negative (support for H4).

Result 1. *Individual capability positively influences task completion.*

Result 2. *Strong competitors negatively influence task completion when feedback is given; it does not influence task completion when strength of competitors is weak.*

After establishing the effects on the mediator task completion, we now turn to the effects on the outcome. The results of ordinary least squares regressions (OLS) are depicted in columns (3) and (4) of Table 2. Task completion has, as expected (H3), a significant positive effect on performance. Workers who complete a task finish more sliders correctly. Capability has a significant direct positive effect on performance (support for H2). People who are capable of doing the task perform better than those who are not. We conclude that the effect of capability on performance is partially mediated by task completion. The more capable a worker is the more likely he will complete the task which will result in a better performance. Furthermore the direct effect of being more capable and as a result perform better does not vanish completely through a mediation of task completion. When feedback is given, a weak competitor has a significant negative effect on performance compared to no feedback. It seems that indeed frontrunners lay back when they know that they are frontrunners. When facing a strong competitor and feedback is given, on the contrary, leads to no different performance than no feedback (regression model 3). The difference between a weak and a strong competitor is significant (regression model 4). As hypothesized, we find a moderating effect of feedback on the effect of strength of competitors on performance. Our hypothesis H5 is, however, only partially supported: as expected, with given feedback, playing against a weak competitor decreases performance; contrary to our expectation, when playing against a strong competitor feedback does not increase performance. These effects are not influenced by fatigue of workers who play longer than those who quit the task, since we control for task completion in our regressions.

Result 3. *Individual capability positively influences performance. The effect is partially mediated by task completion.*

Result 4. *Strength of competitors negatively influences performance. When feedback is given, there is a direct, unmediated negative effect of weak competitors on performance. With strong competitors, the negative effect on performance is fully mediated by task completion.*

5 Conclusions and Further Research

The relationship between financial incentive schemes, performance feedback, and worker performance, has gained new relevance with the omnipresence of online labor markets and crowdsourcing of freelancers. In this paper, we have investigated the relationship between a specific incentive design (ROT) and task performance in an anonymous crowd labor market. We hypothesized a moderated mediation model on the determinants of worker performance in ROTs and tested it experimentally in an online labor market. In ROTs, individual capability positively influences performance; the effect is partially mediated by task completion. We find that strength of competitors negatively influences performance when feedback about the performance is given. For individual workers, feedback that they are performing comparatively well does not affect their tendency to complete the task but tends to reduce their performance. Potentially as feedback signals that the worker does not have to excel to win the competition, or it signals that low performance is the norm, or both. In cases where feedback shows a worker that he is far behind, it increases the tendency of the worker to quit the task. Underlying reasons could be that the worker knows that winning the competition (and hence the financial reward) is unlikely and he cuts his losses in terms of time invested or that he aims to work on tasks where he has a comparative advantage over other workers. Performance of workers who obtain the feedback that they are comparatively weak but who nevertheless continue to work on a task, do not change their effort compared to receiving no feedback.

Overall, the contribution of this paper is threefold: it summarizes the existing evidence on incentives and feedback on performance in a theoretical model, it tests the hypothesized effects experimentally, and it demonstrates the validity in the context of crowd labor markets. The limitations of the present work are straightforward and include the following: First, we only explore strong and weak competitors but no mediocre competitors. Expanding the analysis in this direction might show that moderation of the effect of strength of competitors on performance by feedback is non-linear. Second, with the slider task we explore a rather unnatural setting. In order to increase external validity, the next step is to explore tasks more common to crowd work and to camouflage the experimental context. Furthermore even though we tried to exclude intrinsic motivation as much as possible, it still might play a role for some workers, which should be analyzed in future research. At present, our results have implications for designing feedback schemes in crowd work environments. First off, feedback of a ROT demotivates the lower and upper part of the population. One way to avoid that might be to provide feedback only to selected participants. As a direction for future research, it seems fruitful to develop adaptive feedback systems that provide feedback only if the participant will in expectation act positively on it. In addition, future work should disentangle the effects of social norms and financial incentives on worker performance.

References

Baron, R.M. and D.A. Kenny (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51 (6), 1173-1182.

Berinsky, A.J., G.A. Huber, G.S. Lenz (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. Political Analysis, 20 (3), 351-368.

Bull, C., A. Schotter and K. Weigelt (1987). Tournaments and Piece Rates: An Experimental Study. Journal of Political Economy, 95 (1), 1-33.

Chilton, L.B., J.J. Horton, R.C. Miller and S. Azenkot (2010). Task search in a human computation Market. In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP 2010), pp. 1-9, USA, New York.

Deci, E.L., G. Betley, J. Kahle, L. Abrams and J. Porac (1981). When Trying To Win: Competition and Intrinsic Motivation. Personality and Social Psychology Bulletin, 7 (1), 79-83.

Donabedian, A. (1980). Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to Its Assessment. Volume 1. Health Administration Press, Ann Arbor.

Donabedian, A. (2003). An Introduction to Quality Assurance in Health Care. Oxford University Press, New York.

Eccles, J.S. and A. Wigfield (2002). Motivational Beliefs, Values, and Goals. Annual review of psychology, 53 (1), 109-132.

Eriksson, T., S. Teyssier and M.C. Villeval (2009a). Self-Selection and the Efficiency of Tournaments. Economic Inquiry, 47 (3), 530-548.

Eriksson, T., A. Poulsen and M.C. Villeval (2009b). Feedback and incentives: Experimental evidence. Labour Economics, 16, 679-688.

Fershtman, C. and U. Gneezy (2011). The Tradeoff between Performance and Quitting in High power Tournaments. Journal of the European Economic Association, 9 (2), 318-336.

Fischbacher, U. (2007). Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. Experimental Economics, 10 (2), 171-178.

Gill, D. and V. Prowse (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. American Economic Review, 102 (1), 469-503.

Hayes, A.F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication Monographs, 76 (4), 408-420.

Horton J.J., D.G. Rand and R.J. Zeckhauser (2011). The online laboratory: Conducting experiments in a real labor market. Experimental Economics, 14 (3), 399-425.

Ipeirotis, P.G. (2010). Analyzing the Amazon Mechanical Turk Marketplace. XRDS, 17 (2), 16-21.

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 (HCOMP 2010), pp. 64-67, USA, Washington DC.

Kaufmann, N., T. Schulze and D. Veit (2011). More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In Proceedings of the Seventeenth Americas Conference on Information Systems (AMCIS 2011), USA, Detroit.

Kittur, A., S. Khamkar, P. André and R.E. Kraut (2012). CrowdWeaver: Visually Managing Complex Crowd Work. In Proceedings of the ACM 2012 conference on Computer supported cooperative work (CSCW 2012), pp. 1033-1036, USA, Seattle.

Kittur, A., J.V. Nickerson, M.S. Bernstein, E.M. Gerber, A. Shaw, J. Zimmerman, M. Lease and J.J. Horton (2013). The Future of Crowd Work. In Proceedings of the 2013 conference on Computer supported cooperative work (CSCW 2013), pp. 1301-1318, USA, San Antonio.

Kokkodis, M. and P.G. Ipeirotis (2013). Have you Done Anything Like That? Predicting Performance Using Inter-category Reputation. In Proceedings of the sixth ACM international conference on Web search and data mining (WSDM 2013), pp. 435-444, Italy, Rome.

Kuhnen, C.M. and A. Tymula (2012). Feedback, Self-Esteem, and Performance in Organizations. Management Science, 58 (1), 94-113.

Mason, W. and S. Suri (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods, 44 (1), 1–23.

Mason, W. and D.J. Watts (2009). Financial Incentives and the "Performance of Crowds". ACM SigKDD Explorations Newsletter, 11 (2), 100-108.

Mao, A., Y. Chen, K.Z. Gajos, D. Parkes, A.D. Procaccia and H. Zhang (2012). TurkServer: Enabling Synchronous and Longitudinal Online Experiments. In Proceedings of the HCOMP (2012).

Paolacci, G., J. Chandler and P. Ipeirotis, (2010). Running Experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5 (5), 411–419.

Pilz, D. and H. Gewald (2013). Does Money Matter? Motivational Factors for Participation in Paid- and Non-Profit-Crowdsourcing Communities. In Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013), Germany, Leipzig.

Pull, K.,H. Baker and A. Baker (2013). The Ambivalent Role of Idiosyncratic Risk in Asymmetric Tournaments. Theoretical Economics Letters, 3 (3A), 16-22.

Reeve, J. and E.L. Deci (1996). Elements of the Competitive Situation That Affect Intrinsic Motivation. Personality and Social Psychology Bulletin, 22 (1), 24-33.

Ryan R.M. and E.L. Deci (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary Educational Psychology, 25 (1), 54-67.

Shaw, A.D., J.J. Horton and D.L. Chen (2011). Designing Incentives for Inexpert Human Raters. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW 2011), pp. 275-284, China, Hangzhou.

Teschner, F. and H. Gimpel (2013a). Crowd Labor Markets as Platform for IS Research: First Evidence from Electronic Markets. In Proceedings of the 2013 International Conference on Information Systems (ICIS 2013), Italy, Milan.

Teschner, F. and H. Gimpel (2013b). Validity of MTurk Experiments in IS Research: Results from Electronic Markets. Working Paper.

Van Dijk, F., J. Sonnemans and F. van Winden (2001). Incentive systems in a real effort experiment. European Economic Review, 45 (2), 187-214.

Wang J., P.G. Ipeirotis and F. Provost (2013). Quality-Based Pricing for Crowdsourced Workers. Working Paper. Available at SSRN: http://ssrn.com/abstract=2283000.