



Discussion Paper

How (not) to Incent Crowd Workers

by

Tim Straub¹, Henner Gimpel, Florian Teschner¹, Weinhardt Christof¹

in: Business & Information Systems Engineering, 2015

The final publication is available at:
<http://dx.doi.org/10.1007/s12599-015-0384-2>

¹ Institute of Information Systems and Marketing, Karlsruhe Institute of Technology (KIT)

University of Augsburg, D-86135 Augsburg
Visitors: Universitätsstr. 12, 86159 Augsburg
Phone: +49 821 598-4801 (Fax: -4899)

University of Bayreuth, D-95440 Bayreuth
Visitors: F.-v.-Schiller-Str. 2a, 95444 Bayreuth
Phone: +49 921 55-4710 (Fax: -844710)

WI-499



How (not) to Incent Crowd Workers – Payment Schemes and Feedback in Crowdsourcing

Tim Straub (corresponding author)

*Institute of Information Systems and Marketing
Karlsruhe Service Research Institute
Karlsruhe Institute of Technology (KIT)
tim.straub@kit.edu*

Prof. Dr. Henner Gimpel

*Research Center Finance & Information Management, University of Augsburg
Project Group Business & Information Systems Engineering of Fraunhofer FIT
henner.gimpel@fim-rc.de*

Dr. Florian Teschner

*Institute of Information Systems and Marketing
Karlsruhe Institute of Technology (KIT)
florian.teschner@kit.edu*

Prof. Dr. Christof Weinhardt

*Institute of Information Systems and Marketing
Karlsruhe Service Research Institute
Karlsruhe Institute of Technology (KIT)
christof.weinhardt@kit.edu*

Abstract: Crowdsourcing gains momentum: In digital work places such as Amazon Mechanical Turk, oDesk, Clickworker, 99designs, or InnoCentive it is easy to distribute human work to hundreds or thousands of freelancers. In these crowdsourcing settings, one challenge is to properly incent worker effort to create value. Common incentive schemes are piece rate payments and rank-order tournaments among workers. Tournaments might or might not disclose a worker's current competitive position via a leaderboard. Following an exploratory approach, we derive a model on worker performance in rank-order tournaments and present a series of real effort studies using experimental techniques on an online labor market to test the model and to compare dyadic tournaments to piece rate payments. Data suggests that on average dyadic tournaments do not improve performance compared to a simple piece rate for simple and short crowdsourcing tasks. Furthermore, giving feedback on the competitive position in such tournaments tends to be negatively related to workers' performance. This relation is partially mediated by task completion and moderated by the provision of feedback: When playing against strong competitors, feedback is associated with workers quitting the task altogether and, thus, showing lower performance. When the competitors are weak, workers tend to complete the task but with reduced effort. Overall, individual piece rate payments are most simple to communicate and implement while incenting performance is on par with more complex dyadic tournaments.

Teaser: Tournaments among workers are common: the best gets a price, is promoted or the like. Common sense and theory tell that tournaments create incentives and boost performance beyond what is achieved with individual compensation. For the world's largest crowdsourcing market we empirically show this to be false in certain cases. Not even the best dyadic tournament we designed outperforms simple piece rate payments. For most of simple and short crowd work tasks, employers will be better off using straightforward individual piece rate compensation.

Keywords: *Crowdsourcing, Online labor, Incentives, Exploratory study, Experimental techniques, Real effort task, Rank-order tournament, Piece rate, Feedback.*

Note: Prior versions of some parts of this paper – most notably the research model and preliminary statistical analysis of Study 2 – have been presented as research in progress at the Twenty Second European Conference on Information Systems 2014 and the conference Collective Intelligence 2014 (Straub et al. 2014a; 2014b).

1 Introduction

Recently, crowdsourcing is getting attention by both practitioners and researchers as model to outsource human work on demand to a broad, diverse, and distributed workforce. Paid crowdsourcing today is provided by many commercial vendors, e.g., Amazon Mechanical Turk (MTurk for short), oDesk, Clickworker, 99designs, and InnoCentive. These platforms provide access to a range of different workers who work on a wide range of tasks – from simple repetitive e-mail tagging to creative and more complex tasks such as building logos (Kittur et al. 2012; 2013; Hammon and Hippner 2012).

In these digital labor markets one challenge for organizations is to properly incent worker effort and quality of work. Work relations are short-lived and commonly one-shot labor relations. Quality control is therefore mostly done by repetition of work by different workers (Ipeirotis et al. 2010; Kokkodis and Ipeirotis 2013; Wang et al. 2013). In paid crowdsourcing settings, workers are usually incented by a piece rate (pay per finished task) or by a tournament price. Piece rate payments are most commonly observed for the collection of crowd input on activities that can be divided into small pieces that can be done (mostly) independently of each other (Malone et al. 2010). This type of work can, for example, be data entry, image tagging, or verification of addresses. Paying a price to the best performing crowd worker or a small set of top performers is most commonly seen for contests when only one or a few good solutions are needed. Examples include the design of a good algorithm or logo. On platforms hosting such contests, like 99designs, the employer (i.e. the person or organization that creates a crowdsourcing task and posts it on the crowd labor market) typically has to decide whether to provide feedback on a worker's current competitive position. One can display leaderboards and signal who is the provisional winner or one can hide this information from workers. Rank-order tournaments (ROTs) are also commonly used in traditional work places (Microsoft, GE, Yahoo! etc.) and sports (poker, soccer leagues etc.). Given their wide usage and appeal of using competitive elements to incent workers, some organizations using crowdsourcing even employ ROTs on platforms like MTurk, where piece rate (PR) payments are seen as the standard. Setting up and controlling a ROT is clearly more cumbersome than straightforward PR payments – handling this complexity might, however, pay off when the workers' performance is higher given that both incentive schemes provide the same average wage for crowd workers.

Overall, this raises two important questions: (1) Do rank-order tournaments lead to better crowd worker performance than piece rate payments? (2) When conducting a crowd labor tournament, should one provide feedback on the worker's competitive position? We investigate both these questions in an exploratory way using a series of three real effort studies on MTurk with overall 874 workers participating. The scope of our research are tasks that aim at the collection and later aggregation of crowd input; we do not study settings in which the employer is interested in only a single best solution and our results might not extend to such settings. Furthermore we only analyze short (3 min) dyadic ROTs. Our results might not extend to longitudinal and more complex ROT settings with many participants. For the simple and short task in our study, the surprising results are that providing feedback on workers' competitive position tends to decrease their performance. In a nutshell, the root for this counter-intuitive result is that low performers stop working all together, while high performers knowing that they will be rewarded work less. Thus, the best dyadic ROT is the one without displaying a leaderboard; letting workers in the dark on how likely they are to win. Not even this best dyadic ROT setting we designed does incent higher worker performance than a simple PR payment. Thus – given the simplicity of implementing, communicating, and controlling PR payments – we conclude that PR payments are a better incentive mechanism than a dyadic ROT for short and simple tasks in crowd labor settings.

2 Background and Research Model

2.1 Crowd Work

Crowdsourcing and online labor markets have emerged as new labor pools of human work that allow organizations to flexibly scale their workforce and hire experts, typically for a comparatively low price (Leimeister 2010). Today, MTurk dominates the market for crowdsourcing micro-tasks that are trivial to humans but challenging to computers (Ipeirotis 2010). Recently, experimental researchers increasingly started using MTurk for its relatively low costs and large subject pool. Previous work has examined the validity and costs of MTurk experiments (e.g. Chilton et al. 2010) and worker 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; 2013b) for recent examples.

Two of the main issues with crowd work are (1) how to secure quality and (2) incent workers to give their best (e.g. Wang and Ipeirotis 2013; Shaw et al. 2011). In this paper, we focus on incentives for crowd work. Shaw et al. (2011) show that linking monetary incentives to the responses of other workers (e.g. penalty for disagreeing to the majority) lead to high performance. Paolacci et al. (2010) report that to get comparable results to traditional offline labor settings, crowdsourcing needs rather small monetary incentives. Contrary, Mason and Watts (2009) show that more money leads to more effort, while quality is not affected. Moreover, compared to a piece rate, an overall lower quota pay scheme, which only pays for a set of completed tasks, leads to a greater output. To sum up, it is an open debate which incentive and information structures are best suited to stimulate worker performance.

2.2 Rank-Order Tournaments and Piece Rates

In rank-order tournaments (ROT) two or more people compete against each other and are ranked according to their performance. Only one or more top performer(s) win the tournament. Overall, economics suggest that ROTs incent workers better than piece rates (PRs) (Bracha and Fershtman 2013; Lazear and Rosen 1981; Ehrenberg and Bognanno 1990; Bull et al. 1987; van Dijk et al. 2001). Ehrenberg and Bognanno (1990) find that professional golf players are positively incented by tournaments. Lazear and Rosen (1981) give evidence that ROTs, used in work places, incent risk-averse workers equally well as a PR. Bracha and Fershtman (2013) distinct effort into labor effort and cognitive effort. Participants who work under a ROT exert more labor effort but at the same time less cognitive effort than when working under a PR. Reasons are that competition incents workers but they are less able to do cognitive tasks under pressure. Van Dijk et al. (2001) find that effort levels and variance in ROTs are higher compared to PRs. In addition, low ability workers work harder. Similarly, Bull et al. (1987) find a higher variance of effort in ROTs compared to PRs.

These results suggest the strength of the competitors as explanation for effort variance: Some subjects might lose interest, when falling behind. Others who are in front might relax. Some who are close to each other might actually be competing. Eriksson et al. (2009a) present experimental evidence that if subjects can choose between ROTs and PRs, variance decreases and effort levels increase in ROTs. They further find that risk-averse subjects tend to choose a PR.

Eriksson et al. (2009b) experimentally study the influence on subjects' effort by giving feedback on their current position with PR payments and ROTs. Three different feedback rules on relative performance are observed – no feedback, feedback given half way through the experiment, and a continuously updated feedback. On average feedback does not change effort, but subjects who are behind make more mistakes under continuous feedback and almost never drop out of the ROT. The reason could be a social norm to never give up (Eriksson et al. 2009b). We argue that this relation might, however, be

stronger in a laboratory setting than an anonymous digital crowdsourcing settings. Evidence for this is presented by Fershtman and Gneezy (2011): Quitting is often avoided by participants, because it is socially stigmatized. Still, higher rewards lead subjects to exert more effort and quit more often at the same time. Finally, Pull et al. (2013) show that in dyadic tournaments where ability of subjects is heterogeneous, effort levels should decrease, because both know that one will win anyway. When subjects' abilities are homogeneous, effort levels should be high. In consequence we expect that a continuous feedback will lead to the same effect. In detail, 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). While workers who performed worse than their expectations, will increase their effort but reduce their expectations. This implies that showing feedback has the potential to improve and lower performance of participants depending on the current position in the tournament.

2.4 Research Model

Our first aim is comparing performance of crowd workers in tournaments (ROT) and with piece rate payments (PR). Following Van Dijk et al. (2001), Bracha and Fershtman (2013), and Ehrenberg and Bognanno (1990), we hypothesize that – when both mechanisms yield the same expected payout – ROTs should be associated with higher performance. PR payments offer little room for designing the incentive scheme; the key parameter is the PR itself which is set to be equal to the average ROT payout. A ROT, on the contrary, opens up more design options. To analyze these and aim for the best ROT design, we summarize the related work reviewed above and common sense in a model depicted in Figure 1.

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 be related to the work process and structures. Structural constructs 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 moderated mediation model.

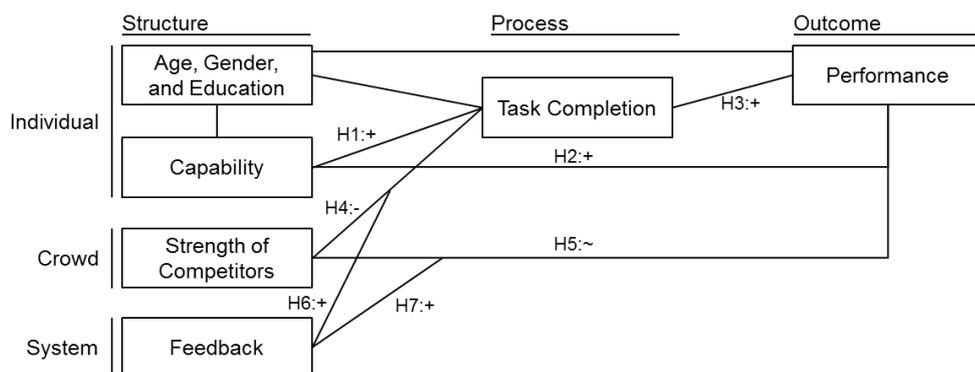


Figure 1. Hypothesized model on the correlates of worker performance in rank-order tournaments

The model shows the hypothesized correlates of crowd worker performance in ROTs. Performance is the achievement of a worker on a given task in a given time frame. We operationalize it as number of successfully completed instances of a task in a certain time frame. Based on common sense we believe that performance is directly related to the worker's capability, i.e. his ability to perform the specific task. Hence, we measure capability as the number of finished tasks in a pre-round. We expect capable workers to perform better. Strength of competitors is the performance of competitors. Feedback indicates whether participants are informed about their current position in the tournament. In our case whether a leaderboard is shown. Based on the work by Eriksson et al. (2009b) we argue that performance might be related to the competitors' strength in cases when

feedback on the performance and current standing in a ROT is provided. Therefore the correlation might be moderated by feedback. Given evidence from studies on ROTs, the direction of the moderated effect of the competitors' strength on performance is, however, not ex-ante clear (Eriksson et al. 2009b; Fershtman and Gneezy 2011; Pull et al., 2013). Whether a worker finishes the task or not is indicated by task completion. Following Fershtman and Gneezy (2011), Eriksson et al. (2009b) and common sense, we assume a strong association with performance. Task completion is hypothesized to mediate the correlation of capability and strength of competitors on performance. Workers able to do a task will finish it more often. Therefore we assume a positive correlation between capability and task completion. Strength of competitors is assumed to be correlated with task completion: Similar to Fershtman and Gneezy (2011), we believe that falling behind leads to quitting, hence, the stronger the competitor, the more likely workers quit the task. Feedback moderates the association of strength of competitors with both task completion and performance. Only when feedback is given the competitors' strength can be seen and hence show a relation. Facing strong competitors is expected to lead to a stronger relation between competitors' strength and task completion than facing weaker competitors (cf. Pull et al. 2013). For strong competitors, we hypothesize the relation to 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 and therefore performance decreases. When he is facing an equally good competitor and always has to excel to win, we expect a performance increase and almost no dropout rates, since he has always a fair chance to win. Finally, we expect a worker's age, gender, and education to be correlated with capability, task completion, and performance – at least these demographics might serve as surrogate measurement for less observable individual characteristics. We do not hypothesize any directions of this correlation, since this is not in focus of this work.

3 Study Design and Procedures

In this paper we explore the relations between performance, strength of competitors, and feedback as summarized in our model. Hence, we present results of three studies: Study 1 compares piece rate payments with the simplest dyadic tournament providing no performance feedback. Study 2 investigates the performance in dyadic tournaments depending on the strength of the competitor and whether feedback is provided or not. Study 3 further tweaks the design of the dyadic tournament by featuring a group matching where individual crowd workers are matched with supposedly equally well performing competitors to spur their performance.

All three studies have similar designs and use experimental procedures.. Experimentation serves different roles in different research traditions. In the information systems literature, Boudreau et al. (2001), for example, posit that experiments take place in settings created by the researcher for the investigation of a phenomenon: the researcher controls independent variables (e.g. feedback), creates different treatment conditions by varying these independent variables, randomly assigns research participants to these treatment conditions, and measures the impact on one or more dependent variables (e.g. performance). Our studies use these experimental techniques. In economics, experimental research has a long tradition. It is accepted that experiments can serve multiple purposes. Roth (1986, 1987), for example, differentiates three classes of experiments under the labels "speaking to theorists", "searching for facts", and "whispering in the ears of princess". Experiments speaking to theorists are designed to test well-articulated formal theories. Experiments searching for facts explore phenomena where existing theory may have little to say; they are "often designed without reference to a specific body of theory" (Roth 1987, p. 2). Experiments whispering in the ears of princess are designed to resemble natural environments and inform policy decisions. On the backdrop of this experimental economics perspective, our exploratory studies are experiments searching for facts (Roth

1986, 1987), more precisely they are framed field experiments (Harrison and List 2004). A contrary perspective common in the social sciences (see e.g. Stebbins 2001) and applied to information systems research by, e.g., Briggs and Schwabe (2011, p. 98) suggests that the goal of experimental research “is to test the propositions of a deductive nomological theory. It may also be called confirmatory research.” In this perspective, only studies “speaking to theorists” (Roth 1986, 1987) can be considered experiments. In order to be clear on the exploratory nature of our research, we refer to our empirical studies as “exploratory studies using experimental techniques”.

In all three studies, we implemented a real effort task, similar to the slider task by Gill and Prowse (2012), to measure worker performance. Workers have a fixed time adjusting as many sliders as possible ranging from 0 to 100 to a value of 50. Correct positioned sliders reset until either the time for the task elapses or the worker quits. The number of sliders a worker correctly sets prior to the end of the task is the measure of performance. The rather simple, needless work is by purpose and typical for real effort experiments. The intention is to measure workers reaction with a task that depends as little as possible on pre-existing knowledge, learning by doing effects, randomness, or guessing (Gill and Prowse 2012). In addition it partially excludes intrinsic motivational factors like entertainment, learning, or contribution to an epic meaning.

All tournaments are dyadic tournaments – a worker competes with only one other worker, the winner gets a bonus of USD 1.00, the loser does not receive a bonus. The choice of the smallest possible number of competitors aims at making the competitor salient and allowing workers to most clearly judge their competitive position. In this, we follow the study design by Eriksson et al. (2009a, 2009b), Fershtman and Gneezy (2011), and van Dijk et al. (2001) and posit that this design feature carves out the relation between tournament competition and performance most clearly. To increase experimental control, participants do not compete live but against historic data collected from a previous participant. This is made clear in the instructions.

All participants are recruited from the general pool of MTurk workers restricted that they can only take part once and in one of the three studies, reside in the US, finished at least 1,000 MTurk tasks (so called HITs) prior to our studies, and 95% of their prior work was approved by the respective employer. Using MTurk as platform for experimental research is gaining prominence in various disciplines, including economics (e.g., Horton et al. 2011), psychology (e.g., Buhrmester et al. 2011), computer science (e.g., Chilton et al. 2010), and information systems (e.g., Teschner and Gimpel 2013a). For the purpose of this study, MTurk is not merely a platform to recruit and reimburse subjects but the natural environment of many crowd workers. In fact it is the crowd labor market with most workers and most tasks. All three studies start out with the instructions and a short quiz to test understanding of these, followed by the experimental task, a questionnaire on some demographics, and payment of participants according to their respective performance.

The studies are conducted with a custom-made web application. From a technical perspective we follow the guidelines of Mao et al. (2012) and Mason and Suri (2012). The slider task was originally developed in z-Tree (Fischbacher 2007). We implemented a similar version to be accessible online through MTurk. An *out* button was added, to allow the workers to quit the task whenever they wanted. Potentially quitting a task is common in crowd labor markets: Considering the experience a worker gains during a task and the opportunity costs of time, it might well be rational for the worker to quit by simply abandoning the task. In the MTurk context this is referred to as not returning a HIT. The explicit option to quit aims at reducing experimenter demand effects and the relevance of a potential social norm to never give up. Figure 2 illustrates the task and the feedback for all three studies: Figure 2a shows an example for a ROT with feedback. At any time during the ROT a 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 be set to 50. In addition, the screen has a timer at the top and a quit button at the

bottom. Figure 2b exemplifies the no feedback treatments; it is identical except that feedback on the competitor's performance is missing – this information is only disclosed after the ROT when the result is shown. The user interface for PR treatments is identical to the one for ROTs without feedback (Figure 2b); the subsequent payment differs.

In statistical tests, we employ a 0.1 level to decide on the rejection of null hypotheses. More detailed information on p-values is provided. Design features that differ between the three studies are described below.

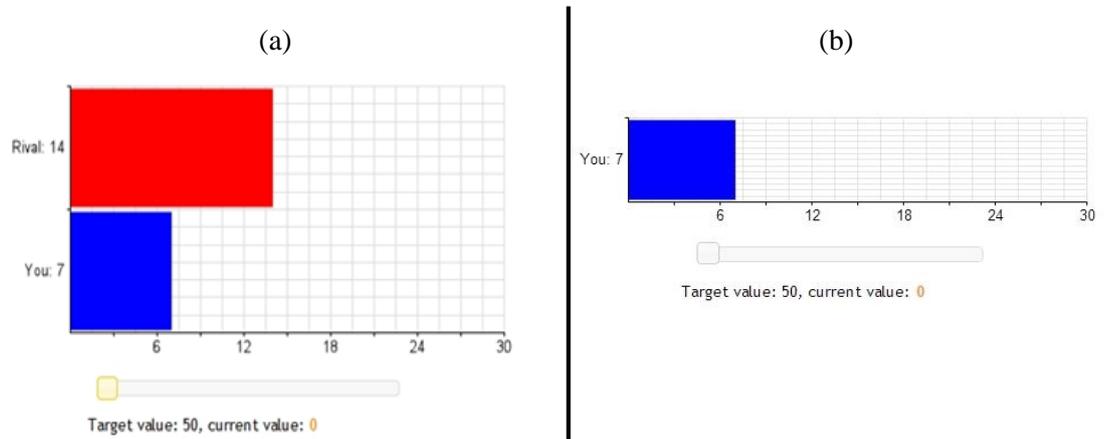


Figure 2. User interface: Feedback (left image), no feedback (right image)

4 Study Results

4.1 Study 1: Piece Rate versus Rank-Order Tournaments

Study 1 is a comparison of piece rate payments (PR) with rank-order tournaments (ROT) providing no feedback on the competitor's performance during the study. Presumably, performance depends on various individual characteristics like the individual capability to perform the task and other factors that might partially be captured by observing age, gender, and education. To account for this partially unobservable heterogeneity, we employ a within-subject comparison for the two treatments (PR and ROT): each subject participates in both payment schemes. Each participant plays a training round of the slider task for 30 seconds to get familiar with the task and the interface followed by two study rounds of two minutes each. One of the study rounds is under PR conditions, getting USD 0.02 per finished slider, while the other round is under ROT conditions, winning USD 1.00 if the participant finished more sliders than his competitor (random tie breaking). Based on pre-tests, payments are calibrated such as participants achieve the same average payment in both mechanisms. Hence, the differences in performance cannot be attributed to different expected or realized payoffs. In both treatments the participants get the same information – their own performance (Figure 2a). For the ROT, they are informed after the round if they won. To control for order effects, wealth effects, learning, fatigue, and the like, we balance the order of the two payment schemes. The number of finished sliders in PR and ROT is the measure for the participants' performance to be compared between payment schemes.

Overall, 149 participants took part in the study. 73 first worked under the PR scheme, then under ROT; 76 first worked under the ROT, then the PR scheme. Participants' age ranges from 19 to 66 years with mean 31 years. 41.6% are female. The task took on average 11 minutes, and the average total payment was USD 1.63. Payment consists of a fix USD 0.50 show-up fee and payments for both incentive schemes. For PR, mean payment was USD 0.55 (SD = 0.17), for ROT it was USD 0.58 (SD = 0.50). Payments in both incentive schemes are statistically indistinguishable (two-sided t-test, $t = -0.571$, $p\text{-value} = 0.569$).

We analyze the relation of the two payment schemes with the participants' performance in three ways: First we count how many participants performed better in PR than in ROT and the other way round. Of 149 participants, 17 finished the exact same number of sliders under both incentive schemes, 63 performed better in the ROT than with PR and 69 performed better with PR than ROT. This data suggests that both incentive schemes are about equal: given one performs differently under PR and ROT, the likelihood of being better under ROT is 48% which is statistically indistinguishable from a random 50% (two-sided binomial test, p-value = 0.664). Second, we compare the mean number of sliders finished in either treatment: 27.64 for PR (SD = 8.26) and 27.65 for ROT (SD = 8.39). Again, no statistically significant difference (two-sided matched pairs t-test, $t = -0.028$, p-value = 0.978). Third, we employ an ordinary least squares (OLS) regression with *performance* as dependent variable (DV) while controlling for age, gender, education and the order effects. The binary variable *round* equals zero for the first incentive scheme and equals one for the second. *Age* is measured in years, *education* in the following categories: some high school completed = 0, high school diploma = 1, some college completed = 2, associate's degree = 3, bachelor's degree = 4, master's degree = 5, doctorate = 6. The binary variable *tournament* is one for ROT and zero for PR. This is the focal variable in this study. The results are depicted in Table 1. Most importantly – but not surprisingly given the other tests described in this paragraph – we do not see a significant correlation between the treatment and performance.

The absence of significance does not directly imply the absence of a relation. Thus it is interesting to analyze the marginal effect size of the tournament in explaining variance in performance. For doing so, we ran a second regression analysis to obtain the residual R^2 , i.e. without *tournament* as independent variable and compare it to the variance explained by *tournament* to calculate the effect size f^2 (Cohen 1988, p. 407ff.) By convention, $f^2 = 0.02$ is termed a small, 0.15 a medium, and 0.35 a large effect. Here, effect size f^2 turns out to be merely 0.00003, i.e. three orders of magnitude less than a small effect. The relation is not only statistically insignificant and meaningless, it is economically meaningless either: the estimated effect of running a tournament equals the estimated effect of increasing the participants' age by about three month which is not substantial given an average age of 31 years. Given the confluence of this evidence, we formulate the following Result 1.

Result 1. *Given equal expected payments, both piece rate and dyadic rank-order tournament payment schemes without feedback on the competitive position result in equal crowd worker performance.*

DV and method	Performance (OLS)
Intercept	33.886 ***
Age	-0.319 ***
Gender male	2.905 **
Education	0.065
Round	3.317 ***
Tournament	0.080
N	149
R^2	0.242

Table 1. Regression results for Study 1 (Significance codes: '***' $p < 0.001$; '**' $p < 0.01$; '*' $p < 0.05$; '+' $p < 0.1$).

Further relevant results from Study 1 are that performance has a strong relation to age (older workers perform worse than younger workers) and gender (males perform better than females), but performance is not related to the education of participants. Thus, in the following studies, we continue to elicit demographic information and use it as control in the analysis. In addition, participants' performance is strongly associated to the order of tasks (participants perform better in the second round). To avoid any confounding effects

from the order of treatments, for the following studies we use a between-subject design and increase sample size to control for individual heterogeneity.

A ROT requires more effort and complexity in implementing, communicating, and controlling than PR payment. As this effort does not translate to higher performance, we conclude that the short and simple dyadic ROT studied in Study 1 is – for practical reasons – less suited than PR payments. This might, however, strongly depend on the ROT’s design, most prominently the lack of feedback on a worker’s current competitive position. Whether such feedback is positively correlated to performance and renders a ROT worthwhile is the focus of Study 2.

4.2 Study 2: Feedback on a Weak, Mediocre or Strong Competitor

Study 2 studies the relation of the strength of competitors and feedback to performance. It is a between-subjects comparison of four treatments. In each treatment workers first work on the slider task for 1.5 minutes with a PR payment of USD 0.01 per finished slider. The number of finished sliders is our measure for capability. In addition, it allows workers to get familiar with the task and interface. We do not use this data to judge whether PR or ROT lead to higher performance. Second, workers participate in a three minute dyadic ROT. For the ROT, each worker is randomized to either of four 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), feedback on the performance of the competitor in a ROT with a mediocre competitor (FM), and feedback on the performance of the competitor in a ROT with a weak competitor (FW). Data for competitors is retrieved from historic data; it is constant for each treatment in order to not induce unnecessary variance. The weak competitor finishes 27 sliders in three minutes time; the mediocre competitor 47 sliders; the strong competitor 66 sliders. The number of sliders a worker finishes in the ROT is the measure for his performance. In case a worker finishes more sliders than his competitor, he wins USD 1.00.

331 workers participated: 97 in NF, 80 in FS, 74 in FM, and 80 in the FW treatment. Participants’ age ranges from 18 to 66 years with mean 32 years. 39.9% are female. The task took on average 8 minutes, and the average payment was USD 0.89.

The moderated mediation model sketched in Figure 1 is evaluated with a set of eight regressions, following the general steps from Hayes (2009) contemporary interpretation of Baron and Kenny’s (1986) mediation and moderation analysis and a bootstrap test of indirect effects following Preacher and Hayes (2004). For the causal step mediation analysis we first establish the correlation of the causal variables on the mediator (regression models 1 to 4) and then estimate the correlation of causal variables and the mediator on the outcomes variable (regression models 5 to 8). *Task Completion* is binary (completed = 1, not completed = 0). *Strength of competitors* is coded in three levels (weak, mediocre, or strong). In our setting, the statistical consideration of moderation differs from the conventional approach: Conventionally, feedback moderating the correlation of strength of competitors would be modeled by two direct effects (one from feedback, one from strength of competitors) and the interaction of these. In our model and experiment, strength of competitors is, however, not meaningfully defined in the absence of feedback. Without feedback, strength of competitors cannot be correlated with either task completion or performance. Thus, moderation here results in four combinations: No feedback (irrespective of the strength of competitors), feedback and a weak competitor, feedback and a mediocre competitor, and feedback and strong competitor. Table 2 provides the results.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DV and method	Task Completion (Logit regression)				Performance (OLS regression)			
Intercept	1.885 +	1.899 +	1.899 +	1.899 +	5.455 +	4.608	4.608	4.608
Age in years	.017	.018	.018	.018	-.146 **	-.140 **	-.140 **	-.140 **

Gender male	-.396	-.314	-.314	-.314	1.326	1.356	1.356	1.356
Education	-.328 *	-.372 *	-.372 *	-.372 *	.103	.164	.164	.164
Task Completion					20.009 ***	20.442 ***	20.442 ***	20.442 ***
Capability	.144 ***	.149 ***	.149 ***	.149 ***	1.236 ***	1.239 ***	1.239 ***	1.239 ***
Feedback	-1.264 *		-.325	-1.438 *	-.980		-2.278 +	-1.188
Weak x Feedback		-.325		1.113 +		-2.278 +		-1.091
Mediocre x Feedback		-1.438 *	-1.113 +			-1.188	1.091	
Strong x Feedback		-1.765 **	-1.440 *	-.327		0.598	2.876 *	1.786
N	331	331	331	331	331	331	331	331
R ²	0.206	0.242	0.242	0.242	0.707	0.711	0.711	0.711

Table 2. Regression results for Study 2 (Significance codes: ‘***’ $p < 0.001$; ‘**’ $p < 0.01$; ‘*’ $p < 0.05$; ‘+’ $p < 0.1$; for logit regressions, Cragg and Uhler’s R²).

As expected, *capability* is substantially associated with *task completion* (regression model 1; support for H1 in the research model). *Feedback* is a dummy equal to 0 for NF treatment and 1 for FW, FM, and FS. The interaction of *strength of competitors* and *feedback* assesses the moderation. When facing a weak competitor and feedback is given there is no significant relation to task completion compared to no feedback. Contrary, when playing against a mediocre or strong competitor, there is a significant relation to task completion. Feedback makes workers quit the task when facing a mediocre or strong competitor. Furthermore, a mediocre or strong competitor makes workers quit more often compared to a weak competitor (regression model 3, significant effect of a mediocre or strong competitor interacted with feedback). In total, feedback moderates the relation of strength of competitors and task completion (support for H6). The stronger the competitor the more likely a worker quits the task resulting in a negative correlation (support for H4).

Result 2. *Individual capability is correlated to task completion in a rank-order tournament. Capable workers finish the task more often.*

Result 3. *Mediocre and strong competitors are correlated to task completion when feedback is given in a rank-order tournament; it is not related to task completion when strength of competitors is weak. Workers quit the task more often when facing stronger competitors.*

After establishing the correlations on the mediator task completion, we now turn to the correlations with the outcome. The results of ordinary least squares regressions (OLS) are depicted in columns (5) to (8) of Table 2. As expected (H3), task completion has a strong relation to performance. Workers who complete a task finish more sliders correctly. Capability has a direct relation with performance (support for H2). Capable workers perform better than those who are not. The correlation of capability on performance is mediated by task completion. The more capable a worker is the more likely he will complete the task which will result in better performance. Giving feedback about a weak competitor is correlated with performance compared to no feedback (regression model 6). Workers who are informed about facing a weak competitor perform worse than without this information. We conclude that indeed frontrunners lay back when they know that they are frontrunners. On the contrary, giving feedback about facing a mediocre or strong competitor leads to no different performance than no feedback (regression model 6). The difference between a weak and a strong competitor is significant (regression model 7). The difference between a weak and a mediocre competitor (regression model 7) and between a mediocre and strong competitor is not significant (regression model 8). As hypothesized, we find a moderating relation of feedback on the relation of strength of competitors on performance. 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 mediocre or strong competitor feedback does not increase performance. These associations are not associated with fatigue of workers who

play longer than those who quit the task, since we control for task completion in our regressions.

Result 4. *Individual capability is related to performance in a rank-order tournament. Capable workers perform better than those who are not. The relation is partially mediated by task completion.*

Result 5. *Strength of competitors is related to performance in a rank-order tournament. When feedback is given, there is a direct, unmediated negative correlation of weak competitors with performance. With mediocre or strong competitors, the negative correlation with performance is mediated by task completion.*

After the causal mediation analysis steps we now turn to the indirect effect and the effect sizes using Preacher and Hayes' (2004) bootstrap test. To do so, our dataset with four treatments is modified in seven sets of pairwise treatment comparisons to access the analysis (Pederson et al. 2011). All results are based on 10,000 bootstrap simulations with a sample size of 331. *Feedback* summarizes treatments FW, FM, and FS. The results are depicted in Table 3.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatments	NF – Feedback	NF – FW	NF – FM	NF – FS	FW – FM	FW – FS	FM – FS
Average mediation effect (95% CI)	-1.940 * [-2.715, -0.132]	-0.248 [-1.731, 1.212]	-1.425 * [-3.591, -0.122]	-2.417 ** [-4.354, -0.608]	-1.301 [-3.507, 0.311]	-2.107 * [-4.234, -0.109]	0.496 [-2.814, 1.640]
Average direct effect (95% CI)	-0.882 [-2.800, 1.074]	-2.218 * [-4.383, -0.051]	-1.057 [-3.625, 1.501]	0.721 [-1.807, 3.278]	1.161 [-1.341, 3.760]	2.939 * [0.471, 5.420]	1.778 [-0.922, 4.424]
Total effect (95% CI)	-2.822 + [-4.729, 0.203]	-2.466 + [-5.195, 0.241]	-2.483 + [-5.833, 0.276]	-1.696 [-4.815, 1.551]	-0.141 [-3.325, 2.627]	0.832 [-2.125, 3.841]	2.274 [-2.191, 4.490]
Proportion mediated	0.687 + [-0.427, 2.451]	0.101 [-1.665, 1.221]	0.574 + [-0.705, 2.759]	1.425 [-8.683, 10.292]	9.258 [-14.144, 15.219]	-2.533 [-22.687, 26.780]	0.218 [-8.562, 9.023]
N	331	331	331	331	331	331	331

Table 3. Mediation analysis results for Study 2 (Significance codes: ‘***’ p < 0.001; ‘**’ p < 0.01; ‘*’ p < 0.05; ‘+’ p < 0.1).

We first compare the NF treatment that resembles the tournament used in Study 1 with all three feedback treatments (FW, FM, FS; model 1). When feedback is given the significant total negative effect on performance is mediated by task completion. In comparison to giving no feedback, feedback on a weak competitor leads to no significant mediation but significant negative direct and total effects on performance (model 2). For a mediocre competitor, the significant negative total effect is mediated by task completion (model 3). For a strong competitor, interestingly, there is a significant negative mediation effect on performance via task completion; the total effect is, however, not significant as the mediation effect is partially offset by an (insignificant) positive direct effect (model 4). Looking at the differences between the different strengths of the competitors (models 5 to 7) only the comparison of the two extremes – FW and FS – shows significant relations. Compared to facing a weak competitor a strong competitor makes some workers quit the task (mediation effect) while incenting others to perform better (direct effect). Both effects cancel each other about out leading to an insignificant total effect. These results further underpin and detail our findings so far: the stronger the competitor, the more likely a worker quits the task. Task completion thereby partially mediates the negative correlation of feedback with performance. When a worker decides to complete the task, the weaker the competitor, the lower is the worker’s performance.

In summary, Study 2 suggests that giving feedback is related to one’s performance. No matter how strong or weak the competitor in a dyadic ROT performs, on average over all workers it decreases performance. The mechanism of this negative correlation is either the mediation by task completion or a direct negative effect on performance. This result seems disillusioning for short dyadic ROTs showing leaderboards. It might, however, be driven by averaging over workers facing competitors of different strength. It still might be the case that a clever matching of workers yields higher performance. Specifically, we hypothesize that a competitor not substantially stronger or weaker but about on par with the worker itself should result in the fiercest competition that does neither discourage continuation nor allow to relax. This issue is addressed by Study 3.

4.3 Study 3: Group Matching

Study 3 – analyzing the relations of an equally well performing competitor – consists again of a PR round measuring capability and a dyadic ROT measuring performance. We compare three treatments for the ROT phase – no feedback about the competitor’s performance (NF3; with a suffix 3 to denote Study 3), feedback about a mediocre competitor (FM3), and feedback about an equally good competitor (FE). The first two exactly replicate the respective treatments from Study 2. FE is new: knowing a worker’s capability from the PR phase, we pick a competitor for the ROT who – in the available historic data – is closest to him in terms of capability.

Overall, 394 workers participated: 131 in NF3, 128 in FM3, and 135 in FE. Participants’ age ranges from 18 to 66 years with mean 34 years. 47.7% are female. The task took on average 11 minutes, and the average payment was USD 1.69. To assess the moderation mediation model we directly use Preacher and Hayes’ (2004) bootstrap test method, resulting in four dichotomous comparisons (NF3 – Feedback, NF3 – FM3, NF3 – FE, FM3 – FE) with each 10.000 bootstrap simulations. Results are summarized in Table 4.

	(1)	(2)	(3)	(4)
Treatments	NF3 – Feedback	NF3 – FM3	NF3 – FE	FM3 – FE
Average mediation effects (95% CI)	-1.219 ** [-2.382, -0.237]	-0.203 + [-2.428, 0.087]	-1.270 * [-2.798, -0.106]	-1.321 [-1.874, 1.245]
Average direct effect (95% CI)	-1.541 * [-2.954, -0.153]	-1.757 * [-3.399, -0.177]	-1.331 [-2.985, 0.349]	0.426 [-1.314, 2.139]
Total effect (95% CI)	-2.761 *** [-4.617, -1.029]	-1.960 ** [-5.028, -0.818]	-2.602 ** [-4.926, -0.595]	-0.896 [-2.247, 2.449]
Proportion mediated (95% CI)	0.442 * [0.126, 0.881]	0.104 + [-0.048, 0.850]	0.488 * [0.061, 1.290]	1.475 [-7.574, 6.444]
N	394	394	394	394

Table 4. Mediation analysis results for Study 3, (Significance codes: ‘***’ p < 0.001; ‘**’ p < 0.01; ‘*’ p < 0.05; ‘+’ p < 0.1).

We first compare the NF3 control treatment with the two feedback treatments (FM3, FE; test 1). We find that feedback has a significant negative total effect on performance. This negative effect is partially mediated by task completion. This reinforces our findings from Study 2 that giving feedback is negatively correlated with worker performance, in some cases through the mediation effect that workers quit the task and in other cases because workers don’t quit but still do less compared to not getting feedback.

The effects of a mediocre competitor on performance observed in Study 2 are replicated (model 2): A mediocre competitor leads to a total negative effect on worker performance with a comparatively low but significant mediation through task completion. The new aspect of Study 3 is studying an equally good – group matched – competitor in treatment FE. Compared to no feedback, this FE leads to a significant negative total effect on performance which is, again, mediated by task completion (model 3). Even though workers have a reasonable chance to win at all times, since their competitor has about equal strength, they still quit the task resulting in lower performance. The correlation seems to be stronger (more negative) than the correlation induced by a mediocre

competitor, but there is not a significant difference between FM3 and FE (model 4). Contrary to our expectations, group matching shows no positive or less negative correlation with performance, but rather a comparable negative correlation. Regarding the implementation overhead, we therefore do not recommend to implement such a matching, since it does not boost workers' performance in a short term dyadic ROT. Reasons for this could be that feedback may just be a distraction or excels arousal.

Result 6. *The existence of an equally good competitor is negatively correlated with performance in rank-order tournaments. When feedback is given, workers facing an equally good competitor perform worse than without feedback. This relation is mediated by task completion.*

5 Conclusions and Future Research

Financial incentive schemes and their relationship to performance feedback and worker performance, have gained new relevance with the omnipresence of digital work places and crowdsourcing human work. In this exploratory study, we have investigated dyadic rank-order tournaments (ROTs) and piece rates (PRs) as incentive schemes for short crowdsourcing tasks and their relationship to task performance in an anonymous digital workplace for activities that can be divided into small pieces that can be done (mostly) independently of each other. We introduced a model on the correlates of worker performance in ROTs and tested it with a series of empirical studies on MTurk – the most popular crowdsourcing workplace. The best dyadic ROT in our studies does not excel a simple PR in terms of performance elicited from participants. Not all dyadic ROTs are equal, however: We find a relation to performance from giving feedback about the competitor's strength. Feedback that a worker is performing comparatively well does not show a relation to his tendency to complete the task but tends to reduce his 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. Feedback that shows that a worker trails behind increases his likelihood to quit the task. Underlying reasons could be that the worker knows that he is about to lose (and hence the financial reward) and he cuts his losses in terms of time invested or he aims to work on tasks where he has a comparative advantage over other workers. Mediocre competitors lead to correlations in between. When competitors are group matched and, hence, compete against an equally strong competitor performance is reduced as well. Reasons could be that workers perform worse under pressure or are distracted by constantly checking the feedback on whether they are winning or not. 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. In summary, this results in a clear guidance how to set up the two studied incentives in an anonymous crowd labor market for distributable work: A simple piece rate payment is better than a short dyadic tournament as incentive for simple short crowdsourcing tasks, as it is easiest to implement and unbeaten in terms of worker performance. Holding a short dyadic contest does not offer performance benefits – if one does anyways, during the tournament one should not provide a leaderboard or feedback on worker's relative performance. Selectively matching workers to homogenous groups seems not to be worth the effort, as it decreases their performance in such a contest setting.

The main contribution of this paper is threefold: First, it summarizes existing evidence of incentives and feedback in tournaments via a theoretical model; Second, it studies the model and compares two common incentive schemes used in crowd work in a series of three studies; Third, it provides guidance for crowdsourcing practitioners on how to set up payment schemes for their crowd workers. It thereby partially answers the practitioner question on how to design crowd labor tasks and contributes to theoretical discussion of designing and developing digital workplaces in general. The limitations of the present work are straightforward and include the following: First, we explore three discrete levels of the strength of a competitor (Study 2) and equally strong competitors (Study 3) but we

do not observe continuous competitor strength. 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, even though the experiment was applied on a crowdsourcing platform (MTurk), the slider task (chosen to provide experimental control on incentives to perform the task) is a rather unnatural task and our short 3-min dyadic tournaments are rather small tournaments. In order to increase external validity even further, a next step might be to explore tasks more common to crowdsourcing, to scale up the tournaments (length and participants), and to camouflage the experimental context. Results might differ when tournaments are played over a longer timeframe with more participants. Third, our feedback system was rather simple. More complex leaderboard and feedback designs might induce different results. Last, we do not look in detail at worker characteristics like, e.g., personality traits that might show, that for some parts of the population, ROTs indeed spur performance.

Future research might investigate causality among the constructs studied in this paper. In our three empirical studies, several constructs depicted in Figure 1 are either controlled by the researchers (strength of competitors, feedback), or they are given exogenously by the nature of participants and vary at most marginally during the short duration of studies (age, gender, education). Pooled with random assignment of participants to treatments, it appears reasonable to hypothesize causation where these constructs correlate with capability, task completion, and performance. Testing for such causation and further investigating the underlying mechanisms is up to future research. In addition, we suggest to extend the analysis to more complex tasks with a longer duration. Other crowdsourcing settings, specifically tasks where the employer is only interested in the single best solution and tasks that require collaboration among crowd workers should be analyzed. Furthermore, it might be fruitful to design tournaments which invoke intrinsic motivation to increase performance. In addition, future work should disentangle the effects of social norms and financial incentives on worker performance.

References

- Baron RM, Kenny DA (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 AJ, Huber GA, Lenz GS (2012) Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3):351-368
- Boudreau M-C, Gefen D, Straub DW (2001) Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly* 25(1):1-16
- Bracha A, Fershtman C (2013) Competitive Incentives: Working Harder or Working Smarter? *Management Science* 59(4):771-781
- Briggs RO, Schwabe G (2011) On Expanding the Scope of Design Science in IS Research. In: Jain H, Sinha AP, Vitharana P (Eds): *DESRIST 2011*, Springer, Berlin/Heidelberg, LNCS 6629, pp. 92-106
- Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6(1):3-5
- Bull C, Schotter A, Weigelt K (1987) Tournaments and Piece Rates: An Experimental Study. *Journal of Political Economy* 95(1):1-33
- Chilton LB, Horton JJ, Miller RC, Azenkot S (2010) Task search in a human computation Market. In: *ACM SIGKDD Workshop on Human Computation (HCOMP 2010)*, USA, New York, pp. 1-9
- Cohen J (1988) *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale
- Donabedian A (1980) *Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to Its Assessment*. Vol 1. Health Administration Press, Ann Arbor.

- Donabedian A (2003) *An Introduction to Quality Assurance in Health Care*. Oxford University Press, New York
- Eccles JS, Wigfield A (2002) Motivational Beliefs, Values, and Goals. *Annual Review of Psychology* 53(1):109-132
- Ehrenberg RG, Bognanno ML (1990) Do Tournaments Have Incentive Effects? *Journal of Political Economy* 98(6):1307-1324
- Eriksson T, Teyssier S, Villeval MC (2009a) Self-Selection and the Efficiency of Tournaments. *Economic Inquiry* 47(3):530-548
- Eriksson T, Poulsen A, Villeval MC (2009b) Feedback and incentives: Experimental evidence. *Labour Economics* 16: 679-688
- Fershtman C, Gneezy U (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, Prowse V (2012) A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review* 102(1):469-503
- Hammon L, Hippner H (2012) Crowdsourcing. *Business & Information Systems Engineering* 4:1-4
- Harrison GW, List JA (2004) Field Experiments. *Journal of Economic Literature* 42(4):1009-1055
- Hayes AF (2009) Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs* 76(4):408-420
- Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14(3):399-425
- Ipeirotis PG (2010) Analyzing the Amazon Mechanical Turk Marketplace. *XRDS* 17(2):16-21
- Ipeirotis PG, Provost F, Wang J (2010) Quality Management on Amazon Mechanical Turk. In: *ACM SIGKDD Workshop on Human Computation (HCOMP 2010)*, USA, Washington DC, pp. 64-67
- Kaufmann N, Schulze T, Veit D (2011) More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In: *Seventeenth Americas Conference on Information Systems (AMCIS 2011)*, USA, Detroit, Paper 340
- Kittur A, Khamkar S, André P, Kraut RE (2012) CrowdWeaver: Visually Managing Complex Crowd Work. In: *ACM 2012 conference on Computer supported cooperative work (CSCW 2012)*, USA, Seattle, pp. 1033-1036
- Kittur A, Nickerson JV, Bernstein MS, Gerber EM, Shaw A, Zimmerman J, Lease M, Horton JJ (2013) The Future of Crowd Work. In: *2013 conference on Computer supported cooperative work (CSCW 2013)*, USA, San Antonio, pp. 1301-1318
- Kokkodis M, Ipeirotis PG (2013) Have you Done Anything Like That? Predicting Performance Using Inter-category Reputation. In: *Sixth ACM international conference on Web search and data mining (WSDM 2013)*, Italy, Rome, pp. 435-444
- Kuhnen CM, Tymula A (2012) Feedback, Self-Esteem, and Performance in Organizations. *Management Science* 58(1):94-113
- Lazear EP, Rosen S (1981) Rank-Order Tournaments as Optimum Labor Contracts. *The Journal of Political Economy* 89(5):841-864
- Leimeister JM (2010) Collective Intelligence. *Business & Information Systems Engineering* 2(4):245-248
- Malone TW, Laubacher R, Dellarocas C (2010) The Collective Intelligence Genome. *MIT Sloan Management Review* 51(3):21-31
- Mason W, Suri S (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1):1-23

- Mason W, Watts DJ (2009) Financial Incentives and the “Performance of Crowds”. *ACM SigKDD Explorations Newsletter* 11(2):100-108
- Mao A, Chen Y, Gajos KZ, Parkes D, Procaccia AD, Zhang H (2012) TurkServer: Enabling Synchronous and Longitudinal Online Experiments. In: *HCOMP (2012)*
- Paolacci G, Chandler J, Ipeirotis PG (2010) Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5):411-419
- Pederson EC, Denson TF, Goss RJ, Vasquez EA, Kelley NJ, Miller, N (2011) The impact of rumination on aggressive thoughts, feelings, arousal, and behavior. *British Journal of Social Psychology* 50:281-301
- Pilz D, Gewald H (2013) Does Money Matter? Motivational Factors for Participation in Paid- and Non-Profit-Crowdsourcing Communities. In: 11th International Conference on Wirtschaftsinformatik (WI2013), Germany, Leipzig
- Preacher KJ, Hayes AF (2004) SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers* 36(4):717-731
- Pull K, Bäker H, Bäker A (2013) The Ambivalent Role of Idiosyncratic Risk in Asymmetric Tournaments. *Theoretical Economics Letters* 3(3A):16-22
- Roth AE (1986) Laboratory Experimentation in Economics. *Economics and Philosophy* 2:245-273
- Roth AE (1987) Introduction and overview. In: Roth AE (Ed) *Laboratory experimentation in economics: Six points of view*. Cambridge University Press, Cambridge, pp. 1-13
- Ryan RM, Deci EL (2000) Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25(1):54-67
- Shaw AD, Horton JJ, Chen DL (2011) Designing Incentives for Inexpert Human Raters. In: *ACM 2011 conference on Computer supported cooperative work (CSCW 2011)*, China, Hangzhou, pp. 275-284
- Stebbins R (2001) *Exploratory Research in the Social Sciences*. Sage Pubs, Thousand Oaks
- Straub T, Gimpel H, Teschner F, Weinhardt C (2014a) Feedback and Performance in Crowd Work: A Real Effort Experiment. *ECIS 2014 Proceedings*, Tel Aviv, Israel.
- Straub T, Gimpel H, Teschner F, (2014b) The Negative Effect of Feedback on Performance in Crowd Labor Tournaments. in: Nickerson, J.; Malone, T. (eds.), *Collective Intelligence 2014: Proceedings*. arXiv.org.
- Teschner F, Gimpel H (2013a) Crowd Labor Markets as Platform for IS Research: First Evidence from Electronic Markets. In: *2013 International Conference on Information Systems (ICIS 2013)*, Italy, Milan
- Teschner F, Gimpel H (2013b) Validity of MTurk Experiments in IS Research: Results from Electronic Markets. Working Paper
- Van Dijk F, Sonnemans J, van Winden F (2001) Incentive systems in a real effort experiment. *European Economic Review* 45(2):187-214
- Wang J, Ipeirotis PG, Provost F (2013) Quality-Based Pricing for Crowdsourced Workers. Working Paper. Available at SSRN: <http://ssrn.com/abstract=2283000>.