CROWD LABOR MARKETS AS PLATFORM FOR IS RESEARCH: FIRST EVIDENCE FROM ELECTRONIC MARKETS

Research-in-Progress

Florian Teschner **Henner Gimpel**

Karlsruhe Institute of Technology (KIT)
Institute of Information Systems and Marketing
Englerstr. 14, 76131 Karlsruhe, Germany

florian.teschner@kit.edu henner.gimpel@kit.edu

Abstract

Crowd labor markets such as Amazon Mechanical Turk (MTurk) have emerged as popular platforms where researchers can inexpensively run web-based experiments. Recent work even suggests that MTurk can be used to run large-scale field experiments such as prediction markets in which participants interact synchronously in real-time. Besides technical issues, several methodological questions arise and lead to the question of how results from MTurk and laboratory experiments compare. In this work we provide first insights into running market experiments on MTurk and compare the key property of markets, information efficiency, to a laboratory setting. The results are mixed at best. On MTurk, information aggregation took place less frequently than in the lab. Our results suggest that MTurk participants cannot handle as much complexity as laboratory participants in time-pressured, synchronized experiments.

Keywords: IS research methodologies, experimental economics, market engineering, market performance

Introduction

In order to closely study group interaction, collaboration and experimental markets, Information System (IS) research typically uses laboratory experiments. With the rise of crowd labor markets such as Amazon Mechanical Turk (MTurk)[1] researchers start to run inexpensively web-based experiments on these platforms. So far the majority of experiments have focused on relatively simple and one-shot interactions with participants (e.g. risk-aversion test, prisoner's dilemma, labeling, etc.). A stream of social science and IS research (e.g. Pilz and Gewald 2013, Kaufmann et al. 2011) shows the validity of running these simple experiments on the platform. Recent work by Mao et al. (2012) suggests that MTurk can be used to run large-scale field synchronized experiments. Advantages include the scale of the subject pool,

[1] https://www.mturk.com/mturk/, accessed on August 30, 2013

comparatively low costs to incentivize subjects, and the ability to conduct experiments without the need for a laboratory infrastructure.

However, these benefits go hand in hand with drawbacks. Firstly, there are technical difficulties of facilitating real-time participant interaction over the web, which is necessary for group decision making or market experiments. Besides these technical issues, several methodological questions arise and lead to the question how results from MTurk and laboratory experiments can be compared. Particularly as Mason and Suri (2010) point out: Are there any conditions in which workers perform differently than in the laboratory setting? These are the question we address in the present paper.

To compare MTurk and lab experiments with real-time participant interaction, we ran prediction market experiments, in which participants aggregate dispersed information via trading. Firstly, we invited US and Indian MTurk workers to participate in an experiment. 110 workers participated. We matched them into cohorts of 3 players and let them play a series of 5 market games. In each game their task was to aggregate dispersed information on two parallel events in an abstract induced preferences setting. For each event, we ran a separate market. Secondly, we replicated a very similar setting in the lab with 54 under-graduate students.

Our preliminary results suggest that running even small-scale prediction markets on MTurk is tough. While participant attention and activity seem to be ok, (behavioral) information aggregation does not occur as commonly and predictably as in the lab and in previous research. Compared to the slightly simplified laboratory setting, the results suggest that MTurk participants cannot handle as much intricacy as laboratory participants in time-pressured, synchronized experiments. Further research is warranted to investigate this issue.

Background & Related Work

Mechanical Turk. Recently Mechanical Turk has gained widespread interest as a platform to run low cost experiments with subjects from a demographically diverse pool. Previous work has documented its validity, costs (e.g. Chilton et al., 2010), and participant demographics (Paolacci et al. 2010). For example Buhrmester et al. (2012) indicate that MTurk participants are undistinguishable from Internet sample on several psychometric scales such as the big five personality traits. More specifically Paolacci et al. (2010) show that workers on MTurk are closer to the U.S. population than subjects from traditional university subject pools. Some recent studies find that results from relatively simple games are consistent with laboratory studies. For example Horton et al. (2011) report similar results for prisoner's dilemma and framing experiments run on MTurk and in the lab. Varying the incentives, Amir et al. (2012) report that MTurk results from dictator, ultimatum, public goods, and trust games are comparable to the lab, even with very low stakes. For a detailed review of research comparing laboratory and MTurk results, please see Mason and Suri (2010).

In an attempt to annul technical difficulties and simplify the process of running synchronized experiments on MTurk, Mao et al. (2012) present a software framework called TurkServer. According to their study, they are able to quickly match 15 participants into cohorts fairly easy.

Given the popularity of Web-based prediction markets like Intrade.com, the Iowa Electronic Market (Berg et al., 2008; http://tippie.uiowa.edu/iem/), or the Economic Indicator Exchange EIX (Teschner et al. 2011) http://www.eix-market.de/), it appears straightforward to conduct prediction market research on MTurk. However, to the best of our knowledge, there have not been any studies running market-based experiments or in particular prediction markets on MTurk.

Prediction Markets. There are various ways to utilize the wisdom or collective intelligence of crowds such as using wikis, reputation systems, or polling mechanisms. Another way to aggregate dispersed information is by setting up a so called prediction market. In these markets, participants trade contracts whose payoff depends on the outcome of uncertain future events. For example, a market contract might reward a dollar if a particular presidential candidate is elected. An individual who thinks the candidate has a 65% chance of being elected should be willing to pay up to 65 cents for such a contract. Market participants form expectations about the outcome of an event. Comparable to financial markets, they buy if they find that prices underestimate probability of the event in question and they sell a stock if they find that prices overestimate the probability of an event. Prediction markets have a long track of successful field applications, e.g., in political elections (Berg et al. 2008), sport events (Luckner et al. 2008), finance (Bennouri et al. 2011), and predicting market development (Spann and Skiera, 2003). See Wolfers and Zitzewitz (2004) and Ledyard et al. (2009) for reviews.

Even though prediction market research is well established there a number of questions that need to be resolved. For example Wolfers and Zitzewitz (2006) highlight five open questions about prediction markets. In the last years research has focused on two types of markets; real-world, anonymous betting markets such as betfair and small scale laboratory markets with induced preferences. Both settings have obvious drawbacks. Running controlled experiments in real markets is impossible as information is not controlled and settings cannot be repeated. On the other extreme, laboratory settings usually limit the number of market participants to a few. Hence, the question arises if results from the laboratory transfer to real markets. MTurk seems to have the potential to run experiments that fill the gap as market experiments can be continuously scaled from a few traders to over a 100 traders.

Moreover, if we are able to run large-scale market experiments on MTurk this hints that other synchronized collaborative or competitive experiments, such as group negotiations, collaborative tasks, or community interactions might be fruitful. Right now all these research directions are limited in the same way as market experiments of limited sample size versus an uncontrolled environment.

Experiment Design

Our experiments study (behavioral) information aggregation in abstract induced preferences setting. The design is a blend of design elements used by Plott and Sunder (1988), Healy et al. (2010), Bennouri et al. (2011), Jian and Sami (2012). There are two binary lotteries represented by two bingo cages, A and B, holding 10 balls each. Some of the balls are black, the others white.

Subjects take the role of experts. They become 'experts' via private information. Once the number of black balls in a bingo cage is randomly determined, each subject receives a private signal that one of the a-priori options (number of black balls) is not true. Subjects' private signals differ. The information structure assures aggregate certainty – would all experts pool their private information, they would know the number of black balls per urn with certainty. However, they cannot communicate directly but exchange information via trading. They rather interaction in prediction markets, specifically in two parallel and identical markets, one for bingo cage A, one for B.[2] The market price is assumed to reflect the aggregate prediction of the probability to draw a black ball from the respective bingo cage. Experts are financially compensated based on their trading performance.

Subjects participate in a series of periods to allow for learning over time and increase the number of observations. In a partner matching, the same cohort of subjects interacts in all periods. Thus, each cohort is independent of any other cohort. For each cohort, periods are independent (except for learning) – bingo cages are newly filled with black and white balls, markets and portfolios are reset.

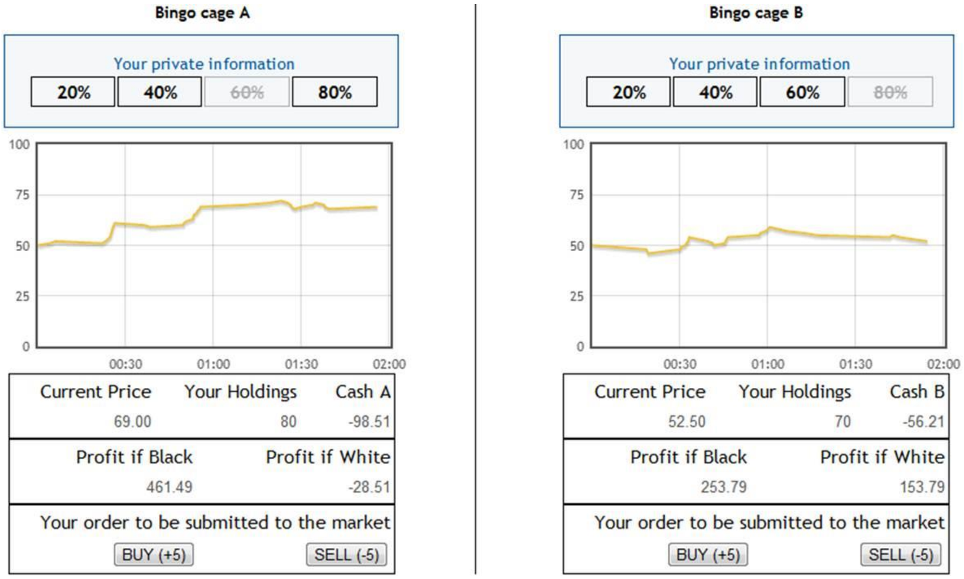
Each period follows 3 phases

1. *Private information and estimate:* The number of black balls per bingo cage is determined and subjects receive private information. Each subject is asked for her private estimate on the probability of drawing a black ball from bingo cage A and from bingo cage B, respectively. Truthful revelation is incentivized with a proper scoring rule (cf. Hanson 2003, p. 109). Private estimates are not communicated to the other experts.
2. *Prediction market:* The market uses a logarithmic market scoring rule (Hanson 2003; Jian and Sami 2012). Subjects can buy and sell virtual stocks for three minutes in two parallel markets (Market A left hand side and Market B right hand side). The value of the stocks is linked to the color of the ball that will be drawn. The final market price is used as the group's best predictor for the number of balls

[2] The lab experiment reported here is part of a larger series of lab experiments. In other settings, it is relevant to have two parallel markets and decide among them. For consistency, we used the same setting with two parallel markets here. As a downside, it increases complexity for subjects. However, having multiple parallel markets is common in most real-world applications of prediction markets.

in the respective bingo cage. Bennouri et al. (2011), Jian and Sami (2012), and others use similar approaches. Figure 1 illustrates the trading interface.

3. *Decision:* A random decision is made to draw a ball from either of the bingo cages (50% chance for each bingo cage). For the bingo cage chosen, the true probability of drawing a black ball is announced to all subjects along with the color of the ball drawn. The true probability determines the value of the shares traded in the prediction market and subjects are paid according to their trading performance. The other market is void; trading in this other market has no impact on subjects' payoff. As the probability for either market being relevant to subjects is 50%, there is no structural reason to favor one or the other market.



Bingo cage A

Your private information

20% | 40% | ~~60%~~ | 80%

Current Price	Your Holdings	Cash A
69.00	80	-98.51

Profit if Black	Profit if White
461.49	-28.51

Your order to be submitted to the market

BUY (+5) SELL (-5)

Bingo cage B

Your private information

20% | 40% | 60% | ~~80%~~

Current Price	Your Holdings	Cash B
52.50	70	-56.21

Profit if Black	Profit if White
253.79	153.79

Your order to be submitted to the market

BUY (+5) SELL (-5)

Figure 1. Trading Screen in the MTurk experiment

After the trading periods, subjects answered a short questionnaire, asking for cognitive reflection (CRT; Fredericks 2005) and demographics.

The MTurk and the lab experiment differ in a few design choices: Most importantly, the subject pool is different with US and Indian MTurk workers on the one hand and German university students on the other hand. On MTurk, cohorts of 3 subjects were matched to interact in the same markets (like Healy et al., 2010). For the lab, we switch to 2 participants per cohort (like Jian and Sami, 2012) for logistics and cost reasons. On MTurk, either 2, 4, 6, or 8 of 10 balls per bingo cage were black. Thus, there were 4 states of nature. Each of 3 experts received the information that one of these states would not be true; leaving a single state that was true. In order to keep the idea of aggregate certainty and mutually exclusive private information with only 2 traders per cohort, we had to reduce the number of states of the world to 3. Thus, in the lab 2, 5, or 8 of 10 balls per bingo cage were black. Finally, both on MTurk and in the lab subjects had a trial period. On MTurk it was followed by 4 payoff-relevant trading periods, in the lab by 10 such periods.

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 (2010). On MTurk, we used opening hours (From 8 AM to 10 PM, CST) to ensure fast partner matching. The understanding of the detailed instructions was tested with nine questions. Also, we used a waiting room for partner matching and properly incentivized participation and performance.

Experiment Results

We start by comparing some descriptive statistics (see Table 1). The laboratory participants are predominantly male and around 22 years old. On MTurk, genders were more equally balanced and participants were older. The differences in duration are due to the different number of rounds (4 vs. 10) played in each setting.

Table 1. Descriptive Statistics

Mean values with variance in brackets

	LAB	MTurk
Gender (female)	26 %	41 %
Age	21.94 (5.39)	32.00 (127.48)
Duration (Trading + Questionnaire)	~39 minutes	~24 minutes
Payment	€12.79 (3.84)	$2.31 (0.21)

The first key question is whether markets aggregated information. The information contained in a forecast can be assessed by regressing actual values on predicted values (Fair-Shiller regressions: Fair and Shiller 1989). In our experiments, actual values are the true probabilities of drawing a black ball. One predictor for this probability is the market price. A second predictor is the average of experts' private estimates. Table 2 displays the results of four Fair-Shiller regressions, one for each market in both settings: The market price contains information on the true probability (β1 is significantly different from zero) in both laboratory markets. On MTurk, the picture is different: Market A aggregates information while Market B does not. In addition, when comparing the estimates for β1 and β2 between the lab and MTurk, coefficients on MTurk are lower for either market.

Table 2. Measuring information contained in market prices

Fair-Shiller regression estimates. Dependent variable: true probability

(Significance code: '*' 0.05)

	LAB Market A	LAB Market B	MTurk Market A	MTurk Market B
α Intercept	-0.359 *	-0.316 *	0.091	0.389
β1 Market price	0.818 *	0.945 *	0.394 *	-0.119
β2 Average estimate[3]	0.791 *	0.621 *	0.283 *	0.339 *
n	270	270	220	220
Adjusted R^2	0.410	0.399	0.071	0.014

To further detail the results, we pool both markets for each setting and run an interaction regression to evaluate differences in the estimates. The dummy variable 'Market B' is unity for market B and zero for market A. Table 3 depicts the results. While in the Lab setting there is no significant interaction effect (estimate β3 and β4) there is a strong effect in the MTurk setting. Hence, we conclude that there is a strong ordering effect on MTurk and none in the lab. In addition, R^2 values in both all model specifications are lower in the MTurk setting. In general MTurk data has been found to be noisy (e.g. Goodman et al. 2013). In our study it seems that MTurk data is indeed noisier than laboratory data.

[3] If both subjects submitted an estimate, the average is calculated. If only one subject provided an estimate, this is taken as aggregate estimate. If neither subject provided an estimate, the observation is dropped from the respective analysis. Alternatives would be to replace missing estimates by the a-priori estimate (50%) or discard these cases from all analyses. Both alternatives lead to the same qualitative results in all statistics.

Table 3. Differences between Market A and Market B
Regression estimates. Dependent variable: true probability
(Significance code: '*' 0.05)

	LAB	MTurk
α Intercept	-0.338 *	0.214 *
β1 Market price	0.800 *	0.379 *
β2 Average estimate	0.771 *	0.044
β3 Market price * Market B	0.164	-0.424 *
β4 Average estimate * Market B	-0.132	0.545 *
n	540	440
Adjusted R^2	0.405	0.078

Result 1: *In line with prior research, laboratory prediction markets aggregate information well. On the contrary, performance of MTurk prediction markets is mediocre.*

The words "well" and "mediocre" in Result 1 are purposefully vague, as evidence from running market experiments on MTurk is still sparse. However, "well" refers to β1 being close to unity (which it would be in case of perfectly informative market prices) for the LAB model in Tables 2 and 3. Jointly with adjusted R^2 values of around 0.4 for the lab, this strongly suggested that substantial information aggregation took place, even if it is not perfect. On the contrary "mediocre" accounts for the fact that information aggregation took place on MTurk (β1 being significantly different from zero for MTurk in Tables 2 and 3). However, information aggregation took place to a lesser degree (comparing β1 estimate and adjusted R^2) than in the lab.

In the lab, subjects played 10 periods. Thus, learning might account for the superior performance of lab markets. However, when restricting the analysis to the first 4 periods of the lab experiment, qualitative differences persist.

Result 2: *Learning cannot explain the differences between lab and MTurk results.*

Subjects on MTurk are in their natural environment, not a sterile lab environment. On MTurk, there is an ample opportunity for distraction like other jobs performed in parallel, web browsing, answering e-mails, chatting, or distractions in the environment. This could explain differences in performance. To assess this, the post-questionnaire featured a multiple-choice question "Were you distracted during the experiment?" 75% of the subjects responded 'not at all', 22% 'a little' and only 3 % 'very'. Thus, while there was some (perceived) distraction, overall subjects were relatively confident in having focused on the experiment. Further research is required to objectify this and compare it to the lab.

We measure subjects' cognitive reflection by posing 3 questions suggested by Frederick (2005). The number of zero to three correct answers is seen as measure for cognitive reflection. Among lab subjects, the average score was 2.11 compared to 1.51 for MTurk subjects. The difference is significant (Mann-Whitney-U, p-value: <1%).

Considering only the MTurk results, the difference of markets A and B is puzzling. Both markets are virtually identical: The underlying asset and information are the same, market mechanism, interface, traders, timing etc. are all identical. There is only a difference in names ("A" and "B") and ordering. In the instructions, market A is always named first. On the trading screen, market A is on the left, market B on the right (Figure 1). Objectively, there should be no difference. And in the lab there is no difference. Why do markets perform differently on MTurk?

One could hypothesize, that – given cognitive reflection and focus on the experimental task – the experiment was more cognitively demanding for MTurk subjects than lab subjects. If so, traders might focus their attention more strongly on the 'first' market A, leading to more activity and better information aggregation in A than B. However, when analyzing trading activity and participant attention, we find no differences (t-stat: 1.14; p-value: 0.24) in trading activity between markets A and B in the MTurk setting (see Table 4). The overall activity was even higher when compared to the laboratory setting, certainly partially due to more traders on MTurk. The results hold when we look at the time between the first and the last trade in each market. Participants in both the laboratory and MTurk seem to evenly split their attention and direct the same effort to both markets. Hence we cannot explain why information aggregation takes place in market A and does not in market B. Our best guess is that MTurk participants do not cope as well as laboratory participants with the intricacy of trading in two parallel markets.

Conjecture: *Differences in cognitive reflection and focus on the experimental task might explain differences between lab and MTurk experiments.*

Table 4. Trading activity and participant attention

	LAB Market A	LAB Market B	MTurk Market A	MTurk Market B
Number of trades per cohort and market	33.37	33.48	45.36	41.68
Time between first and last trade (seconds)	94.77	90.14	106.63	102.33

Discussion & Conclusion

Recent work by Mao et al. (2012) suggested a novel way of running low cost synchronized experiments with an attractive subject pool Mechanical Turk. In this work we provide first insights into running market experiments on MTurk and compare the key property of markets – information efficiency – to a similar laboratory setting. The results so far are mixed at best. Information aggregation took place in only half of our markets on MTurk. One reason for this might be the intricacy which participants are willing or able to handle. Moreover, the collected data on MTurk is noisier than data from the laboratory.

From our result 1, we conclude that we cannot expect that all laboratory results can be as well attained on MTurk. This might be especially true for complex settings. Also, we would like to highlight that researchers have to invest even more time and effort compared to laboratory settings into streamlining participant recruitment, preparing instructions and trial periods as well as experiment implementation to successfully run these experiments. This is especially true for a simple, intuitive interface which needs to meet the worker expectations.

The limitations of the present work are straightforward: Most importantly, the laboratory experiment setting has only two participants per market while in the MTurk setting we matched three participants. Thus running the laboratory setting on MTurk or vice versa is the next step.

References

Bennouri, M., Gimpel, H. and Robert, J. "Measuring the impact of information aggregation mechanisms: An experimental investigation," *Journal of Economic Behavior & Organization* (78:3), 2011, pp. 302–318.

Buhrmester, M., Kwang, T. and Gosling, S. D. "Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data?," *Perspectives on Psychological Science* (6:1), 2011, pp. 3–5.

Berg, J. E., Nelson, F. D. and Rietz, T. A. "Prediction market accuracy in the long run," *International Journal of Forecasting* (24:2), 2008, pp. 285–300.

Berg, J. E. and Rietz, T. A. "Prediction Markets as Decision Support Systems," *Information Systems Frontiers* (5:1), 2003, pp. 79–93.

Chen, Y., Kash, I., Ruberry, M. and Shnayder, V. "Decision markets with good incentives," *Internet and Network Economics*, Springer, 2011, pp. 72–83.

Chilton, L. B., Horton, J. J., Miller, R. C. and Azenkot, S. "Task search in a human computation market," *Proceedings of the ACM SIGKDD workshop on human computation*, ACM, 2010, pp. 1–9.

Fair, R. C. and Shiller, R. J. "The informational context of ex-ante forecasts," *The Review of Economics and Statistics* (71), 1989, pp. 325–331.

Frederick, S. "Cognitive reflection and decision making," *The Journal of Economic Perspectives* (19:4), 2005, pp. 25–42.

Goodman, J. K., Cryder, C. E. and Cheema, A. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making* (26:3), 2013, 213–224.

Hanson, R. "Combinatorial Information Market Design," *Information Systems Frontiers* (5:1), 2003, pp. 107–119.

Hanson, R. "Decision Markets for Policy Advice", *in* Patashnik, E. and Gerber, A., ed., *Promoting the General Welfare: American Democracy and the Political Economy of Government Performance*, Brookings Institution Press, Washington D.C., 2006.

Hanson, R., Oprea, R. and Porter, D. "Information Aggregation and Manipulation in an Experimental Market," *Journal of Economic Behavior and Organization* (60:4), 2006, pp. 449–459.

Healy, P. J., Linardi, S., Lowery, J. R. and Ledyard, J. O. "Prediction Markets: Alternative Mechanisms for Complex Environments with Few Traders," *Management Science* (56:11), 2010, pp. 1977–1996.

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

Jian, L. and Sami, R. "Aggregation and Manipulation in Prediction Markets: Effects of Trading ;echanism and Information Distribution," *Management Science* (58:1), 2012, pp. 123–140.

Kaufmann, N., Schulze, T. and Veit, D. "More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk," *Proceedings of the Seventeenth Americas Conference on Information Systems*, 2011, pp. 1–11.

Mao, A., Chen, Y., Gajos, K. Z., Parkes, D., Procaccia, A. D. and Zhang, H. "Turkserver: Enabling synchronous and longitudinal online experiments," *Proceedings of HCOMP* (12), 2012.

Mason, W. and Suri, S. "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior research methods* (44:1), 2012, pp. 1–23.

Oliven, K. and Rietz, T. A. "Suckers Are Born but Markets Are Made: Individual Rationality, Arbitrage, and Market Efficiency on an Electronic Futures Market," *Management Science* (50:3), 2004, pp. 336–351.

Othman, A. and Sandholm, T. "Decision rules and decision markets," *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 625–632.

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

Pilz, D. and Gewald, H. "Does Money Matter? Motivational Factors for Participation in Paid-and Non-Profit-Crowdsourcing Communities," *Wirtschaftsinformatik Proceedings*. Paper 37, 2013.

Plott, C. R. and Sunder, S. "Rational expectations and the aggregation of diverse information in laboratory security markets," *Econometrica: Journal of the Econometric Society* (56:5), 1988, pp. 1085–1118.

Teschner, F., Mazarakis, A., Riordan, R. and Weinhardt, C. "Participation, Feedback & Incentives in a Competitive Forecasting Community," *Proceedings of the International Conference on Information Systems (ICIS*), Shanghai, China, Paper 16, 2011, pp. 1–14

Wolfers, J., and Zitzewitz, E. Five open questions about prediction markets (No. w12060). National Bureau of Economic Research. 2006