Article

Social Psychological and Personality Science
1-9

DOI: 10.1177/1948550619837002
journals.sagepub.com/home/spp


Turkers of the World Unite: Multilevel In-Group Bias Among Crowdworkers on Amazon Mechanical Turk

Abdullah Almaatouq[1], Peter Krafft[1], Yarrow Dunham[2], David G. Rand[1], and Alex Pentland[1]

Abstract

Crowdsourcing has become an indispensable tool in the behavioral sciences. Often, the "crowd" is considered a black box for gathering impersonal but generalizable data. Researchers sometimes seem to forget that crowdworkers are people with social contexts, unique personalities, and lives. To test this possibility, we measure how crowdworkers ($N = 2{,}337$, preregistered) share a monetary endowment in a Dictator Game with another Mechanical Turk (MTurk) worker, a worker from another crowdworking platform, or a randomly selected stranger. Results indicate preferential in-group treatment for MTurk workers in particular and for crowdworkers in general. Cooperation levels from typical anonymous economic games on MTurk are not a good proxy for anonymous interactions and may generalize most readily only to the intragroup context.

Keywords

in-group, dictator game, economic games, Amazon Mechanical Turk, crowdworkers, behavioral economics

Amazon Mechanical Turk (MTurk) is an online platform for crowdsourced labor used extensively by behavioral and computer scientists to conduct scientific experiments and test products. Requesters on MTurk design "microtasks" called human intelligence tasks (HITs) and post requests for crowdworkers to complete them for prespecified wages. Amazon presents MTurk as a kind of impersonal black box for microtask labor, and requesters therefore naturally conceive of the platform as such. Consequently, it is easy to assume that MTurk is a kind of idealized "frictionless" world of "spherical chickens" (Stellman, 1973), where the responses are free from any effects of the social contexts.

The Use of Online Labor Markets

The availability of a large and cheap labor market has had a large impact on human subjects' research. Crowdsourced labor from MTurk is now widely used for obtaining large-scale training data for machine learning systems (e.g., image labeling: Gebru, Krause, Deng, & Fei-Fei, 2017; Russakovsky et al., 2015; text analysis: Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012) and to elicit "human evaluations" as baselines for automated algorithms (Morris, Inkpen, & Venolia, 2014; Zhou, Cummins, Lalmas, & Jose, 2013). It is also increasingly popular for conducting behavioral studies. For example, there have been efforts to systematically replicate classic results from the social sciences (Arechar, Gächter, & Molleman, 2017; Berinsky, Huber, & Lenz, 2012; Chesney, Chuah, & Hoffmann, 2009; Hergueux & Jacquemet, 2015; Horton, Rand, & Zeckhauser, 2011), which in many cases appear to be as reliable as data obtained via traditional methods.

MTurk has also been widely adopted within social psychology. The use of crowdworkers has had a profound influence on the nature and pace of data collection and has opened new avenues to cost-effective replication and extension of familiar research paradigms. This shift has had special impact in areas such as studies of cooperation and conflict, person perception, intergroup attitudes and stereotypes, and group behavior, where cumbersome interactive multiparticipant experiment can be conducted much more easily via online platforms (Hawkins, 2015). At the same time, academics have also engaged in some hand-wringing about this major shift in data collection strategies, and academic studies of the MTurk community itself have increased in quantity and scope. Some of this work has begun to peer behind the veil of crowdworking platforms, highlighting the social context of the people engaged in this form of

[1] Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Yale University, New Haven, CT, USA

Corresponding Author:
Abdullah Almaatouq, Massachusetts Institute of Technology, 77 Mass Ave., Cambridge, MA 02139, USA.
Email: amaatouq@mit.edu

labor and emphasizing that Turkers are not cogs in a distributed human computer or a psychological data simulator. For example, there is evidence that learning over time affects results across behavioral experiments conducted on MTurk (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Rand et al., 2014) and that Turkers sometimes collaborate on tasks that are assumed to be independent (Gray, Suri, Ali, & Kulkarni, 2016; Yin, Gray, Suri, & Vaughan, 2016).

In the present work, we explore the possibility that another central aspect of human behavior "intrudes" on the setting within which MTurk studies occur, namely social identification as a crowdworker. That is, MTurk crowdworkers' sense of community and the "Turker" identity may affect the results of social psychological studies conducted on MTurk involving, for example, cooperation, collaboration, or group dynamics. Why should we suspect that many crowdworkers on MTurk have a sense of community identity associated with their work? First and foremost, as demonstrated by a long history of work in the social identity tradition (Hewstone, Rubin, & Willis, 2002; Hogg, 2016), human psychology is powerfully oriented toward the pursuit of group identity. In addition to the familiar social identities springing up around culturally salient social groups such as ethnicity and nationality, humans are predisposed to attach themselves to newly encountered an otherwise meaningless groups, including groups based on unfamiliar properties such as over- or underestimating dot arrays (Tajfel, 1970) or even groups that are randomly assigned (Billig & Tajfel, 1973). This favoritism occurs along many dimensions and goes far beyond explicit judgments, in that membership in such groups can also affect more subtle implicit attitudes as well as a wide range of behavioral outcomes, including those associated with cooperation and generosity (for a recent review, see Dunham, 2018). What's more, these forms of in-group favoritism can occur even outside an explicit intergroup context, apparently engendered by the mere sense that one is collaborating with or otherwise aligned with others (L. Gaertner, Iuzzini, Witt, & Oriña, 2006). Returning the MTurk context, while it is often assumed that MTurk workers are performing as isolated and self-interested cogs, there is actually a direct reason to think that things are not so simple. Most directly, MTurk workers actively engage in communication and information sharing that might be thought to foster precisely these forms of emergent social identities.

Many online forums exist for MTurk crowdworkers, such as Turker Nation, the /r/mturk subreddit, or MTurkGrind, and interaction on such forums could foster a sense of shared purpose and social identity. A hint of such a collective identity comes from ethnographic analysis of publicly available content on Turker Nation, a forum for MTurk users (Martin, Hanrahan, O'Neill, & Gupta, 2014). This study revealed that crowdworkers use such forums to share well-paying work, discuss employers, educate newcomers, collaborate in doing tasks, provide social support, and consult with employers. Similar results were also revealed by an ethnographic study lasting 19 months involving over 100 crowdworkers (Gray et al., 2016). Their key finding was that crowdworkers do indeed collaborate with each other, often to make up for technical or social shortcomings in the platform. Their work inspired further inquiry by (Yin et al., 2016) to quantitatively investigate the structure and scale of the overall communication network on MTurk. Data from more than 10,000 crowdworkers showed that forums, in particular, play a key role in allowing crowdworkers to communicate. The presence of rich network structure within a single crowdworker platform undercuts claims to independence, as workers clearly are engaged in active communication and sometimes collaboration with one another, forces that are plausibly sufficient to induce collective identity, at least of the minimal extent necessary to foster in-group biases.

There have also been significant efforts specifically aimed at fostering collective action by organizing the labor force of MTurk. For example, Turkopticon is a platform created by researchers (Irani & Silberman, 2013) that provides a way for crowdworkers to rate requesters, empowering crowdworkers to reject low-paying or otherwise exploitative work. Dynamo (Salehi, Irani, Bernstein, & Alkhatib, 2015) is another platform developed by researchers to improve Turkers' capacity for collective action.

Of course, this evidence is indirect; it is possible that only a small selection of the crowdworkers on MTurk use such services and that most Turkers use the platform in isolation. Further, the fact that many crowdworkers rely on MTurk for income might impede the emergence of collective identity. However, even if this is the case, the psychology of social identity provides ample reason to think that even quite minimal associations—such as the most basic shared sense of being a Turker—could induce a motivation toward social affiliation. Groups consist of sets of individuals who share some aspect of their identities, that is, some aspect of their sense of self. Shared identity, in turn, elicits in-group bias, that is, a tendency to favor fellow group members, and is one of the most consistent and reliable findings in the social sciences (Hewstone et al., 2002). For example, it is widely observed with respect to salient real-world social groupings such as ethnicity (Whitt & Wilson, 2007), religiosity (Tan & Vogel, 2008), and political affiliation (Rand et al., 2009). Critically, however, in-group bias also readily emerges based on trivial real-world social groupings (e.g., caused by Pokémon Go teams; Peysakhovich & Rand, 2017) and can be artificially constructed in the laboratory (Brewer, 1979; Dunham, 2018), including when group assignments are explicitly random (Diehl, 1990; Dunham, 2013). Given that in-group bias can emerge under such minimal conditions, it seems plausible that it might occur between workers on the same crowdworker platform or even between anyone who identifies—even in a minimal way—as a crowdworker. If such bias does occur, it would raise important constraints on the generalizability of findings from MTurk, especially in contexts in which in-group bias might operate. We expect bias from these effects to be most severe in one of the areas where MTurk has been disproportionately convenient. In such contexts, where participants are embedded with each other into a shared environment, we expect their behavior to be most affected by their sense of shared identity since their

actions, attitudes, or presence will be directly visible to each other. In our study, we directly tested this possibility.

The Present Study

We conducted a behavioral experiment on MTurk in order to assess the degree of in-group bias among the crowdworkers there. The premise of our experimental design is to use the Dictator Game (DG) to compare how willing crowdworkers on MTurk are to cooperate with each other, with crowdworkers from a different platform, or with random strangers. We elected to use a DG because it reflects consequential real-world behavior (i.e., giving money), and while it is frequently conceptualized as a measure of generosity, in intergroup contexts it directly relates to social identification (Misch, Fergusson, & Dunham, 2018; Peysakhovich & Rand, 2017) and so can serve as a behavioral indicator of group alignment. To determine a realistic floor in giving, we also include an additional condition in which "donated" money is destroyed, that is, lost to all parties including the donating participant. Participants decided how to share a sum of money—a US$1 bonus—between themselves and an anonymous counterparty. We refer to the amount shared as the "donation" of the active participant and examine donation rates across our four conditions (other MTurk worker, nonMTurk crowdworker, stranger, and the condition in which donated money is destroyed). We also elicited verbal explanations for decisions, giving us a qualitative window into the factors motivating allocations.

To further amplify the logic of our inquiry, in many MTurk studies, workers are paired with another worker as a proxy for an anonymous interaction with a stranger. If that assumption holds, then our first three conditions should result in similar levels of donation. On the other hand, if crowdworkers hold a tacit social identity favoring other MTurk workers in particular, or other crowdworkers in general, then we would expect those conditions to result in larger donations than donations to strangers. To forecast our results, we find robust evidence of in-group bias by MTurk crowdworkers. Donations to other crowdworkers on MTurk were 15% higher than those given to crowdworkers from another platform (i.e., CrowdFlower) and 35% higher than those given to a randomly selected person from around the world. Our results thus confirm that being a "Turker" is a strong enough identity to have a sizable impact on experiments conducted on MTurk.

Method

Participants

We recruited 2,500 participants (our preregistered sample size) on MTurk by posting a HIT for the experiment, entitled "Make a decision and complete a short survey," a neutral title that was accurate without disclosing the purpose of the experiment. We excluded 163 participants (i.e., 6.5%) from the study due to failing a comprehension check question, described below. Descriptive statistics providing basic demographic characteristics of the 2,337 valid participants are shown in Figure 1.

Procedure

Each MTurk crowdworker was randomly assigned to one of the four experimental conditions, described below. Each participant was paired with only one recipient (depending on treatment condition) and did not know that other conditions existed. This design minimizes the possibility of an experimenter demand effect for in-group bias.

Consistent with standard payment rates on MTurk, participants received a show-up fee of US$0.20 (for an approximately two-min HIT duration) and then had the opportunity to earn an additional bonus of up to US$1 based on their decision in the study. After making their decision, participants completed a survey with a comprehension check question and a set of measures and demographic questions that might be relevant to their decision including gender, age, education level, income level, membership in other crowdwork platforms (i.e., how many other crowdwork platform they use), prior experience on MTurk (i.e., number of years), the reason for making their decision, and the expected decision made by other participants in this study. The comprehension check asked the participants to enter how much money they would receive as their bonus for this HIT, based on their decision (where the correct answer is: US$1 minus amount donated, see below). In the interest of brevity, we present the postsurvey questions in Online Appendix A.

The crux of our design involves comparing giving in the DG across four randomly assigned treatment conditions involving slightly different messages to participants:

MTurk partner (hereafter *MTurk*): "In this HIT, you can receive a bonus of up to US$1, or you can choose to split part of that money off, with the remainder going to a randomly selected Mechanical Turk crowdworker. How much of the US$1.00 bonus would you like to give to this other Mechanical Turk crowdworker?"

Crowdworker from another platform partner (hereafter *Crowdworker*): "In this HIT, you can receive a bonus of up to US$1, or you can choose to split part of that money off, with the remainder going to a randomly chosen crowdworker from the CrowdFlower microwork platform. How much of the US$1.00 bonus would you like to give to this CrowdFlower crowdworker?"

*Non-crowdworker partner (*hereafter *Random)*: "In this HIT, you can receive a bonus of up to US$1, or you can choose to split part of that money off, with the remainder going to a randomly chosen person in the world (selected based on a randomly chosen postal address). How much of the US$1.00 bonus would you like to give to this randomly selected person?"

Destroying the bonus (hereafter *Destroyed)*: "In this HIT, you can receive a bonus, or you can choose to split part of that money off, with the remainder to be destroyed. How much of the US$1.00 bonus would you like to be destroyed?"

We predicted that donations to MTurk partners would be greater than donations to other crowdworker partners and to

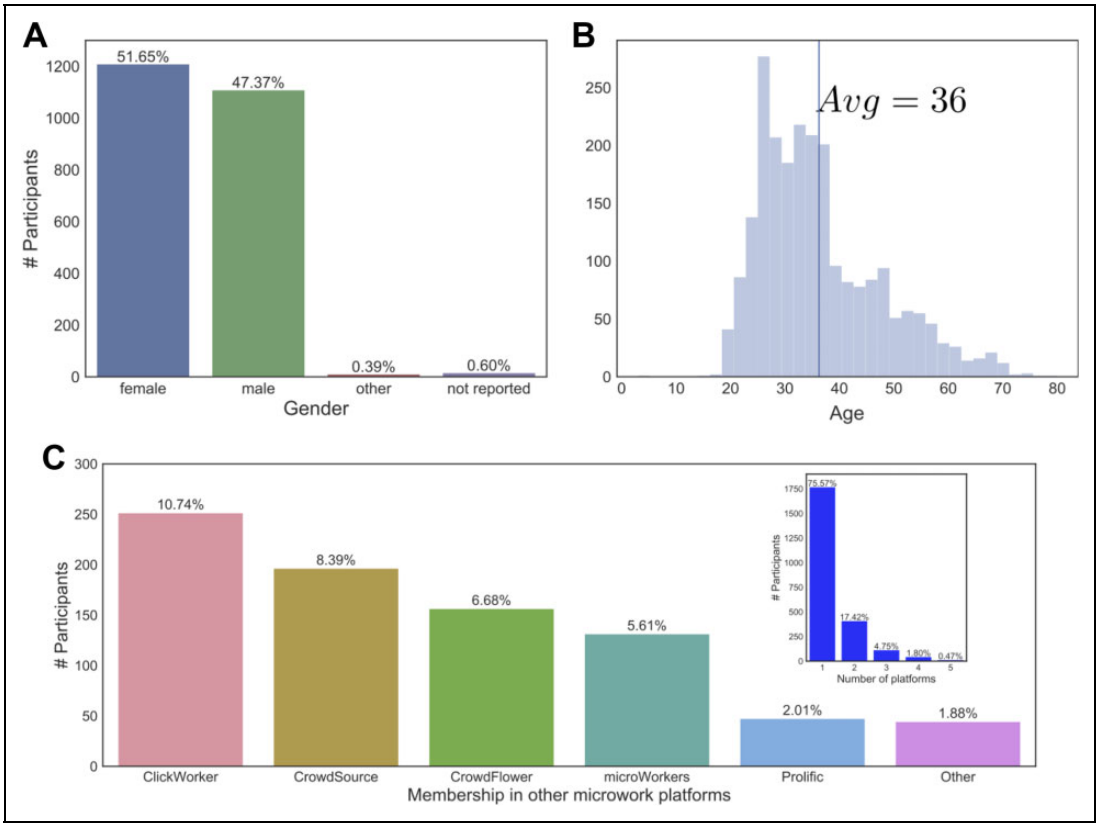

Figure 1. Breakdowns for the age, gender, and memberships in other crowdwork platforms for our recruited participants. The inset figure shows the number of participants that reported being workers on other platforms, where 1 refers to "only being on MTurk."

anonymous partners (and greater than destroyed "donations"). This would indicate the preferential treatment of MTurk crowdworkers over other identity categories. Moreover, if a more general crowdworker identity is present, it would imply that *Crowdworker* > *Random* in terms of the amount donated. Finally, *Random* > *Destroyed* means that participants would rather share the bonus with a complete stranger than destroy it.

We use nonparametric Mann–Whitney tests for our main analysis rather than *t* tests following standard practice for economic games in the experimental economics literature, as the distribution of behavior in these games is typically strongly non-normal (see Online Appendix B). Our main hypotheses, experimental design, and analyses were preregistered before the collection of the data.1 We, therefore, report one-tailed tests for preregistered directional hypotheses. In this work, we report all measures, manipulations, and exclusions. In Online Appendix C, we also perform one-tailed parametric *t* tests for our main results as a robustness check and report sensitivity power analyses that describe the minimum effect size that could be obtained given our study's design.

Measures

Our measure of interest was a DG, a one-person decision process in which the player, "the dictator," determines (or "dictates") how much, if any, of an endowment (US$1 bonus) to donate to a counterpart. The counterpart, "the recipient," simply receives the donation from the dictator. Based on pure self-interest, the dictator should keep all the bonus, donating nothing to the recipient. However, considerable research finds that participants generally donate a nontrivial amount in a wide variety of experimental conditions (Engel, 2011) and that DG giving is related to other measures of cooperation in both economic game and nongame contexts (Peysakhovich, Nowak, & Rand, 2014). Therefore, giving anything in the DG is considered prosocial (sometimes referred to as behavioral "altruism"). The DG has been widely used in behavioral economics and psychological studies on cooperation, altruism, and in-group bias (Johnson & Mislin, 2011; Lane, 2016), making it an appropriate test case to explore our primary research question.

Results and Discussion

Quantitative Analysis

As predicted, one-tailed Mann–Whitney indicates that MTurk crowdworkers donated significantly more as a percentage of their endowment to other MTurk crowdworkers ($M_{MTurk} = 16\%$) compared to crowdworkers from another platform ($M_{Crowdworker} = 14\%$), $U = 235{,}887.0$, $p = .027$. They also donated more to both types of crowdworkers ($M_{Mturk} = 16\%$, $M_{Crowdworker} = 14\%$) than to random persons ($M_{Random} = 12\%$), likely noncrowdworkers, $U = 220{,}531.5$, $p < 10^{-4}$,

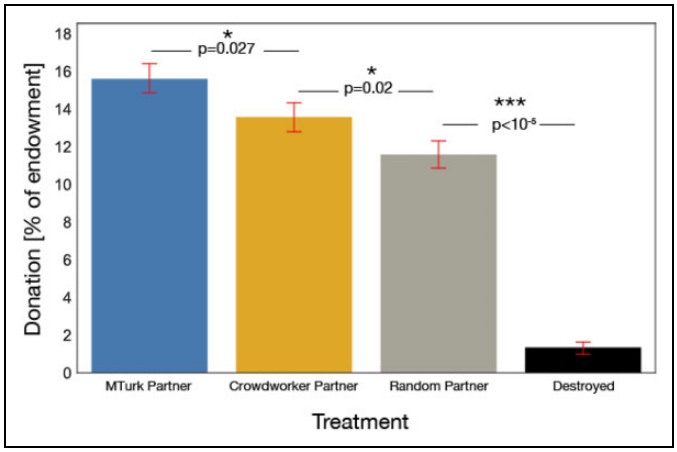

Figure 2. Significant in-group bias exists among Amazon's mechanical Turk crowdworkers (MTurk > Crowdworker; $p = .027$). Also, a significant in-group bias exists for crowdworkers in general (*Crowdworker > Random*; $p = .019$). Finally, there is a preference to donate the bonus to a random person over destroying it (*Random > Destroyed*; $p < 10^{-10}$). *p* Values reported and displayed in the figure were determined by one-tailed Mann–Whitney U tests (our preregistered test) as donation amounts are not normally distributed. Error bars indicate standard error of the mean.

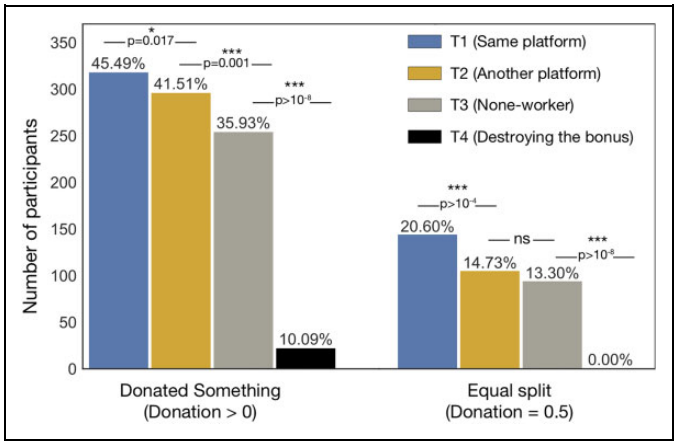

Figure 3. The proportion of participants who donated anything (donation > 0) or made an equal split (donation = 0.5) is greater when the recipient is another Amazon's Mechanical Turk crowdworker compared to when the recipient is a crowdworker from another platform or a random person (likely noncrowdworker). The difference between conditions in hypergenerous donations (donation > 0.5) was not significant (not shown in the figure). *p* Values reported were determined by one-tailed proportions z tests. The number at the top of each bar represents the percentage of participants in each condition that made such a decision.

and $U = 238{,}020.0$, $p = .019$, respectively (see Figure 2). Relatively, the donation to crowdworkers from the same platform was 15% higher on average than the donation given to crowdworkers from another platform and 35% higher than the donation given to a random (likely) noncrowdworker. Also, MTurk crowdworkers significantly preferred to donate to a random person ($M_{Random} = 12\%$) over destroying part of the bonus ($M_{Destroyed} = 1\%$), $U = 55{,}559.0$, $p < 10^{-10}$. The amount of donations destroyed was close to zero, indicating that participants were taking the study seriously and not responding randomly.

Although we observed greater than zero average donation across conditions, most participants did not donate anything (i.e., median donation = 0). Therefore, we further investigate the in-group/out-group differences by computing the proportion of participants that gave anything as a donation (i.e., donation > 0), made an equal split (i.e., donation = 0.5), or made a hypergenerous split (i.e., donation > 0.5; Figure 3). Using one-tailed proportion *z* test, we found that the proportion of participants who donated anything is significantly greater when the recipient is another MTurk crowdworker (45.5%) compared to when the recipient is a crowdworker from another platform (41.5%), $z = 2.1$, $p = .017$, or a random person (36%), $z = 5.0$, $p < .001$. And again, when the recipient is a crowdworker from another platform, the proportion of participants that donated anything is higher (45.5%) compared to when the recipient is a random stranger (36%), $z = 3$, $p = .001$. The effect of in-group bias is even larger when we look at the proportion of participants who made an equal split. In particular, the proportion of participants who made an equal split is much greater when the recipient is another MTurk crowdworker (20.6%) compared to a crowdworker from another platform (14.7%), $z = 3.8$, $p < .001$, or a random stranger (13.3%), $z = 4.8$, $p < .001$. The difference in equal splits between crowdworker from another platform (14.7%) and a random stranger (13.3%) is not significant, $z = 1.1$, $p = .14$. There were no equal splits in *Destroyed* (i.e., destroying the bonus). We explore potential moderators of donation patterns (i.e., heterogeneity in treatment effects) in Online Appendix D. We found that less experienced MTurk crowdworkers donated more overall. These trends are also confirmed when the variables are represented as continuous rather than discretized in our regression results in Online Appendix E. This is consistent with prior work (Capraro, Jordan, & Rand, 2014; Rand et al., 2014), which has consistently shown that more experienced MTurkers are less cooperative.

Qualitative Analysis

Our quantitative results show that MTurk crowdworkers were more generous to other MTurk crowdworkers than to crowdworkers from another platform (CrowdFlower) or to random strangers. While we hypothesized that this resulted from an emergent if a subtle sense of collective identity, in order to explore this possibility in more detail, we conducted a qualitative analysis of the responses from our postexperiment surveys in which participants indicated the reason they made their donation decision.

To conduct our qualitative analysis systematically, we tagged approximately 40% of the reasons that participants provided for giving and approximately 25% of the reasons for not giving. We use this sample from the population of all survey responses to assess the proportions of participants that provided each of the common reasons associated with our codes. Reasons for giving included (i) altruism (i.e., feelings of generosity

Table 1. Percentage of Participants Reporting Each Commonly Provided Reason for Their Decision.

	Treatments			
Reasons Reported	MTurk (%)	CrowdFlower (%)	Random Stranger (%)	Destroying the Bonus (%)
To give				
Altruism	22	18	20	0
Reciprocity	3	1	0	0
Fairness	14	13	8	0
Not to Give				
Self-interest	45	48	44	87
Out-group	6	11	20	0
Fairness	5	5	5	0
Reciprocity	6	5	2	0

Note. We omit the counts of misunderstandings from our tables for clarity–they generally did not vary across conditions. MTurk = Amazon Mechanical Turk

or empathy), (ii) reciprocity (i.e., expectation that it will be reciprocated either by the same interaction partner or a different individual), and (iii) fairness (i.e., citing fairness directly as a reason to give without mentioning an expectation of reciprocity). On the other hand, reasons for not giving included (i) self-interest (i.e., the need for extra income), (ii) out-group (i.e., mentioning out-group as a reason not to give), (iii) fairness (i.e., since they were the ones actually doing the work of the task and not the recipient), and (iv) reciprocity (i.e., citing an expectation that the other would not give). We provide direct quotes from the workers in Online Appendix F.

Counting the frequency of these reasons across experimental conditions provides some insight into the mechanisms behind our quantitative results (see Table 1). We find that participants were slightly more prone to considerations of reciprocity as both a reason to give and a reason not to give to other MTurk workers. Altruism did not vary much between conditions (besides *Destroyed*), but participants may have been slightly less altruistic with CrowdFlower workers. Participants tended to try to be fair with any type of crowdworker but didn't consider fairness so much with random strangers. Levels of self-interested considerations also did not vary much between conditions (again, besides *Destroyed*). The most interesting source of variance we observe is in the proportion of participants citing out-group considerations as a reason not to give. Many more participants mentioned not knowing what the recipient was like or similar reasons in the random stranger and CrowdFlower conditions than when giving to fellow MTurk workers. This prominent difference between conditions in our count of out-group reasoning further supports our argument that cooperation levels may be inflated by in-group bias.

General Discussion

The main contribution of this work is showing that even if crowdworkers do not communicate directly, the fact that they belong to a coherent community of MTurk crowdworkers produces in-group bias that can affect the results of experiments conducted on the platform. Our findings are meant to draw attention to the inevitable contextual factors at play in the social contexts of online labor markets. Keeping these contextual factors in mind can aid in the interpretation of scientific experiments and user studies conducted on such platforms and especially in interpreting the generalizability of findings from these enterprises.

Interestingly, the pattern of data we observe suggests that crowdworkers adopt both a superordinate identity as crowdworkers (hence the greater donations to crowdworkers than strangers) and a subordinate identity as an MTurk worker (hence the greater donations to MTurk workers than CrowdFlower workers). These dynamics are familiar to studies focusing on other intergroup domains, in which superordinate identities can serve as a means of creating group cohesion and reducing in-group bias (S. L. Gaertner, Dovidio, Nier, Ward, & Banker, 1999), but we know of little past work demonstrating how readily they spring up even in an online labor market like MTurk.

One puzzling observation we made was experienced workers show a smaller rather than larger in-group effect. One possibility is that workers could be more excited about the MTurk community when they first join but later become either disillusioned or treat it as more a part of their daily grind. Another possibility, supported by our qualitative analysis, is that experienced workers have played similar games in the past and been burned. For example, two respondents stated, "past experiences where people didn't share" and "in my experience, most people keep the bonus for themselves so if I were another player I wouldn't get anything either." In these cases, workers may be learning to specifically not share with other workers on the platform, whereas their identity as humans and willingness to share with other people (random strangers) still could remain.

We acknowledge that many effect sizes we observed were generally small, including the difference in mean donation between conditions; however, other effects were larger. The probability of participants donating anything at all was 20 percentage points higher with fellow MTurk workers than with random strangers, which is a substantial fraction. These varying effect sizes suggest that the reliability of MTurk as a platform for experiments that generate generalizable conclusions depends on the research question being asked. For example, if a researcher is interested in the overall proportion of individuals who engage in prosocial giving with anonymous others in a task like the DG, our results suggest that a standard MTurk study in which participants are paired with other MTurk crowdworkers will substantially overestimate the rate of truly anonymous giving. Other designs, even the *stranger* condition we employed here, might provide a better estimate. Of course, this insight may also generalize beyond crowdwork platforms. Another common participant recruitment strategy is undergraduate pools, generally students in an undergraduate psychology course. If participants drawn from such a pool are paired with one another, our findings again suggest that social identification with other students will affect results even if no identity contrast is explicitly present.

Another consideration is that the key comparison between MTurk partner and random partner may depend on operational characteristics of these conditions. For example, different procedures to select and donate to strangers might lead to different cooperation levels. If it was possible to bonus nonMTurk strangers in the same way as MTurk participants, would we observe similar cooperation levels? While we find it somewhat unlikely, we cannot rule out the possibility that a different system of administering payments underlies the difference between the MTurk and random condition. Most critically though, even if one thought the difference between MTurk and random conditions depended on method of disbursement, we also show a difference between MTurk and CrowdFlower, meaning that in a more tightly controlled case with two very similar platforms we still show that tacit assumptions about the identity of the recipient affect rates of generosity. Some participants were confused about our treatments. At least a handful of participants did not recognize that CrowdFlower was a crowdwork platform, thinking instead that it was a crowdfunding organization or represented some type of charitable cause. Other participants thought we were deceiving them and did not believe we would actually bonus random workers or send money to random people in the world. However, our estimates indicate that the rates of these misunderstandings did not vary substantially across conditions, with only about 4–5% of participants expressing some such misunderstanding, with these proportions being statistically indistinguishable across experimental conditions (for further details see Online Appendix F).

On the flip side, a reader may also be concerned that the significant in-group bias effects we observe may be an artifact of the fact that our manipulations make shared identity highly salient. The primary purpose of our study is to understand how to interpret existing work in areas such as cooperation research that conducts many of their behavioral experiments using MTurk as a platform. Our main methodological need to accomplish this goal is to have the degree to which identity is made to be a salient characteristic match the degree in existing studies. Accordingly, examining the first 15 papers presenting MTurk DG paper returned by a Google Scholar search for "amazon mechanical turk dictator game" reveals that the vast majority (13 of the 15) papers explicitly stated that the game partner was an MTurker, as in our setup.

A related concern regards MTurk as a source of convenience samples. By definition, studies of convenience samples cannot provide generalizable evidence about the prevalence of a certain phenomenon (e.g., cooperation; Horton et al., 2011). The ability of MTurk to provide generalizable evidence about psychological phenomena is therefore an empirical question, and one that has received considerable empirical attention of late, with the evidence suggesting that data from MTurk in fact usually do approximate in-lab data collection (Casler, Bickel, & Hackett, 2013) as well as data from other online sources (Buhrmester, Kwang, & Gosling, 2011); further, in the area most centrally at interest here, research on cooperation, results from MTurk also appear similar to those collected from more traditional sources (Amir, Rand, & Gal, 2012). That said, the point of our paper is to identify a specific threat to generalizability that is endemic to common uses of the platform and perhaps is shared in alternative participant pools. In other words, we would remind the reader that a large share of current research in social psychology uses MTurk as the basis for general inferences about human psychology, and in our paper, we call attention to a way in which such inferences can systematically go awry.

Conclusion

We have demonstrated that in-group bias exists in MTurk crowdworkers, both toward other crowdworkers in general, but even more powerfully when the recipient is also from MTurk. The critical implication is that levels of cooperation may be higher between participants interacting with each other on MTurk than between random strangers. Thus, more caution should be taken when interpreting, and most importantly generalizing from, studies conducted on MTurk, especially in domains in which in-group bias might most plausibly occur, including studies of cooperation and conflict, person perception, and intergroup attitudes and stereotypes. Our results also indicate that, although the standard way in which cooperation research is conducted on MTurk inflates cooperation, there are simple ways to rephrase instructions that decrease in-group bias.

Authors' Note

All participants who provided explicit consent to participants in this study and Committee on the Use of Humans as Experimental subjects (COUHES) approved the consent procedure. All data collected in the experiment could be associated only with participant's Amazon crowdworker ID on MTurk, not with any personally identifiable information. All participants remain anonymous for the entire study. The study (Approval#: 1509172301) was reviewed and approved by the COUHES at MIT. The data are publicly available on Abdullah Almaatouq; Peter Krafft; Yarrow Dunham; David G. Rand; Alex (Sandy) Pentland, 2018, "Replication Data for: Turkers of the World Unite: Multilevel In-Group Bias Amongst Crowdworkers On Amazon Mechanical Turk," doi:10.7910/DVN/LLLOVE, Harvard Dataverse, V1. Our main hypotheses, experimental design, and analyses were preregistered before the collection of the data. Anonymized preregistration link can be found at http://aspredicted.org/blind.php?x=3dj5q5

Acknowledgments

We would like to thank Matthew J. Salganik, Iyad Rahwan, and Christopher A. Bail for useful comments and discussions.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The experiment in this article was sponsored by Russell Sage Foundation as part of the Summer Institute in Computational Social Science 2017.

ORCID iD

Abdullah Almaatouq https://orcid.org/0000-0002-8467-9123

Supplemental Material

The supplemental material is available in the online version of the article.

Note

1. Anonymized preregistration link http://aspredicted.org/blind.php?x=3dj5q5

References

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the Internet: The effect of $1 stakes. *PLoS One*, *7*, e31461. doi:10.1371/journal.pone.0031461

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the Internet: The effect of $1 stakes. *PLoS One*, *7*, e31461.

Arechar, A. A., Gächter, S., & Molleman, L. (2017). Conducting interactive experiments online. *Experimental Economics*, *21*, 1–33.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*. Retrieved from https://academic.oup.com/pan/article-abstract/20/3/351/1514250

Billig, M., & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, *3*, 27–52.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*. Retrieved from http://psycnet.apa.org/journals/bul/86/2/307/

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *PsycEXTRA Dataset*. doi:10.1037/e527772014-223

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi:10.1177/1745691610393980

Capraro, V., Jordan, J. J., & Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports*, *4*, 6790.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156–2160. doi:10.1016/j.chb.2013.05.009

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*, 1131–1139.

Chesney, T., Chuah, S. H., & Hoffmann, R. (2009). Virtual world experimentation: An exploratory study. *Journal of Economic Behavior & Organization*, *72*, 618–635.

Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings. *European Review of Social Psychology*, *1*, 263–292.

Dunham, Y. (2013). Balanced identity in the minimal groups paradigm. *PLoS One*, *8*, e84205.

Dunham, Y. (2018). Mere membership. Retrieved from https://psyarxiv.com/ux2g9/

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*, 583–610.

Gaertner, L., Iuzzini, J., Witt, M. G., & Oriña, M. M. (2006). Us without them: Evidence for an intragroup origin of positive in-group regard. *Journal of Personality and Social Psychology*, *90*, 426–439.

Gaertner, S. L., Dovidio, J. F., Nier, J. A., Ward, C. M., & Banker, B. S. (1999). Across cultural divides: The value of a superordinate identity. *Russell Sage Foundation*. Retrieved from http://psycnet.apa.org/psycinfo/1999-02753-005

Gebru, T., Krause, J., Deng, J., & Fei-Fei, L. (2017). Scalable annotation of fine-grained categories without experts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1877–1881). New York, NY: ACM.

Gray, M. L., Suri, S., Ali, S. S., & Kulkarni, D. (2016). The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing—CSCW '16*. doi:10.1145/2818048.2819942

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*, 966–976.

Hergueux, J., & Jacquemet, N. (2015). Social preferences in the online laboratory: A randomized experiment. *Experimental Economics*, *18*, 251–283.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*, 575–604.

Hogg, M. A. (2016). Social identity theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding peace and conflict through social identity theory: contemporary global perspectives* (pp. 3–17). Cham, Switzerland: Springer.

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

Irani, L. C., & Silberman, M. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. [*Proceedings of the International ACM SIGCHI Conference on Supporting Group Work*. ACM SIGCHI International Conference on Supporting Group Work. Retrieved from http://dl.acm.org/citation.cfm?id=2470742. Proceedings of the International ACM SIGCHI Conference on Supporting Group Work. ACM SIGCHI International Conference on Supporting Group Work."]

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*, 865–889.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, *90*, 375–402.

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a Turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 224–235). New York, NY: ACM.

Misch, A., Fergusson, G., & Dunham, Y. (2018). Temporal Dynamics of Partisan Identity Fusion and Prosociality during the 2016 US Presidential Election. Retrieved from https://psyarxiv.com/bhxwp/

Morris, M. R., Inkpen, K., & Venolia, G. (2014). Remote shopping advice: Enhancing in-store shopping with social technologies. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 662–673). New York, NY: ACM.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a "cooperative phenotype" that is domain general and temporally stable. *Nature Communications*, *5*, 4939.

Peysakhovich, A., & Rand, D. G. (2017). In-group favoritism caused by Pokémon GO and the use of machine learning to learn its mechanisms. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2908978

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677.

Rand, D. G., Pfeiffer, T., Dreber, A., Sheketoff, R. W., Wernerfelt, N. C., & Benkler, Y. (2009). Dynamic remodeling of in-group bias during the 2008 presidential election. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6187–6191.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Salehi, N., Irani, L. C., Bernstein, M. S., & Alkhatib, A. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd*. Retrieved from http://dl.acm.org/citation.cfm?id=2702508

Stellman, S. D. (1973). A spherical chicken. *Science*, *182*, 1296.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*, 96–103.

Tan, J. H. W., & Vogel, C. (2008). Religion and trust: An experimental study. *Journal of Economic Psychology*, *29*, 832–848.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 U.S. Presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115–120). Stroudsburg, PA: Association for Computational Linguistics.

Whitt, S., & Wilson, R. K. (2007). The dictator game, fairness and ethnicity in Postwar Bosnia. *American Journal of Political Science*, *51*, 655–668.

Yin, M., Gray, M. L., Suri, S., & Vaughan, J. W. (2016). The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web—WWW '16*. doi:10.1145/2872427.2883036

Zhou, K., Cummins, R., Lalmas, M., & Jose, J. M. (2013). Which vertical search engines are relevant? In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 1557–1568). New York, NY: ACM.

Author Biographies

Abdullah Almaatouq is a PhD candidate at MIT and a research assistant at the Human Dynamics group. His research interests lie in the area where computational and social science meet.

Peter Krafft is a Moore/Sloan & WRF Innovation in Data Science Postdoctoral Fellow at the University of Washington Information School and a part-time postdoctoral researcher at the University of California Berkeley Social Science Matrix. His research focuses on collective behavior and particularly the successes and failures of collective intelligence. His research combines experimental methods, observational data analysis, and qualitative approaches.

Yarrow Dunham is an assistant professor of psychology and cognitive science and the director of the Social Cognitive Development Lab at socialcogdev.com. His research focuses on studying how knowledge of social groups is acquired, both in cognitively mature adults and in the developing children. His research combines a range of experimental and cross-cultural methodologies.

David G. Rand is an associate professor of management science and brain and cognitive sciences at MIT and the director of the Human Cooperation Laboratory and the Applied Cooperation Team. His research combines a range of theoretical and experimental methods in an effort to explain the high levels of cooperation that typify human societies.

Alex Pentland is a faculty at MIT and leads both the Human Dynamics and Connection Science research groups. He is a member of the U.S. National Academies, and his most recent books are Social Physics (Penguin) and Honest Signals (MIT Press).

Handling Editor: Gregory Webster